Label-Free Optical Imaging of Chromophores and Genome Analysis at the Single Cell Level

Dissertation Advisor: Professor Xiaoliang Sunney Xie Sijia Lu

Label-Free Optical Imaging of Chromophores and Genome Analysis At the Single Cell Level

Abstract

Since the emergence of biology as a quantitative science in the past century, a lot of biological discoveries have been driven by milestone technical advances such as X-ray crystallography, fluorescence microscopy and high-throughput sequencing. Fluorescence microscopy is widely used to explore the nanoscale cellular world because of its superb sensitivity and spatial resolution. However, many species (e.g. lipids, small proteins) are non-fluorescent and are difficult to label without disturbing their native functions. In the first part of the dissertation, we explore using three different contrast mechanisms for label-free imaging of these species – absorption and stimulated emission (Chapter 2), heat generation and diffusion (Chapter 3) and nonlinear scattering (Chapter 4). We demonstrate label-free imaging of blood vessels, cytochromes, drugs for photodynamic therapy, and muscle and brain tissues with three dimensional optical sectioning capability.

With the rapid development of high throughput genotyping techniques, genome analysis is currently routinely done genome-wide with single nucleotide resolution. However, a large amount of starting materials are often required for whole genome analysis. The dynamic

changes in DNA molecules generate intra-sample heterogeneity. Even with the same genome content, different cells often have very different transcriptome profiles in a functional organism. Such intra-sample heterogeneities in the genome and transcriptome are often masked by ensemble analysis. In this second part of the dissertation, we first introduce a whole genome amplification method with high coverage in sequencing single human cells (Chapter 6). We then use the technique to study meiotic recombinations in sperm cells from an individual (Chapter 7). We further develop a technique that enables digital counting of genome fragments and whole genome haplotyping in single cells (Chapter 8). And we introduce our ongoing efforts on single cell transcriptome analysis (Chapter 9). In the end, we introduce our initial effort in exploring the genome accessibility at the single cell level (Chapter 9). Through the development of techniques probing the single cell genome, transcriptome and possibly epigenome, we hope to provide a toolbox for studying biological processes with genome-wide and single cell resolution.

Contents

Acknowledgements

I am deeply indebted to many individuals who have guided me, accompanied me and stimulated me. I could have never done the work described in this dissertation without the help and encouragement from Prof. Sunney Xie. I am deeply impressed by Sunney's courage and determination of taking on important and challenging projects, through which I built up the self-confidence and discovered the potential of myself. Sunney's courage and attitude of tackling big challenges is contagious and will be remembered and cherished throughout my professional and personal life.

The first part of this dissertation was developed in close collaboration with Dr. Wei Min. Wei has been a great mentor and friend, his enthusiasm in science has stimulated lots of heated late-night discussions in the lab. His sharpness in thinking and his great sense of tackling problems have been a tremendous treasure and I am grateful for the joint adventure we have had.

I am grateful to Dr. Chenghang Zong for his persistence with the challenging project, which led to the critical technical advancement described in Chapter 6 and provides a foundation for all the following chapters. At one point, the project seemed impossibly difficult. I am grateful for the experience of fighting through the challenges together with Chenghang.

I am particularly lucky to have the chance of working with an extremely talented graduate student Alec Chapman. The second part of the dissertation would not be possible

without his contributions in data analysis. I am deeply impressed by his insightful thoughts and understandings on the projects and I have learned a lot through daily discussions with him. I thank Alec for offering a 'minicourse' on bioinformatics to our sequencing people, which I found particularly helpful.

I am grateful to Shasha Chong, Dr. Markus Rueckel and Dr. Gary Holtom for their contributions on the projects mentioned in Chapters 2 to 4, and I thank Drs. Christian Freudiger and Brian Saar for discussions. I particularly want to thank Gary for handling the laser systems extremely well. Without Gary, the projects would have taken us at least two more years, if at all possible.

I thank Jenny Lu, for her contributions in developing the experiments mentioned in Chapters 5, 7 and 8. Jenny is the smartest and most diligent undergraduate student I met and I am grateful for the chance of working with her for almost two years. I thank Zi He, for his contribution in developing single cell transcriptome. We have gone through ups and downs together on a daily basis and I treasure this experience. I particularly need to thank Dr. Song Lin, for taking care of the lab, and for feeding us with great food at 4pm everyday.

During my time in the Xie lab, I have been fortunate to interact with many great students and scientists who I have not been working directly with. I thank Drs. Katsuyuki Shiroguchi, Peter Sims, Will Greenleaf, David Sutter, Steve Mao, Huiyi Chen, Chongyi Chen and Jun Yong for daily interactions and discussions on the projects. I particularly need to thank Katsu and Peter. Katsu and I discussed almost everyday and he helped me a lot in developing the projects in the right direction and in developing my critical thinking. I have not had many

conversations with Peter, but I have been inspired and stimulated each time I talked to Peter, and I am very grateful for that.

For me, working in the Xie group was not about fighting projects by myself. It has been a special experience that great students and scientists worked together and help each other, on projects that are fundamental in science and important for technical advancement. I am grateful to everybody in the group and for the special experience I have had during my five years in the group.

I am grateful for the chance of working in BIOPIC at Peking University on a collaborative project (Chapter 6) in the summer of 2010. It was a very delightful experience and I interacted with a lot of intelligent individuals. I thank Prof. Fuchou Tang, Prof. Yanyi Huang, Prof. Ruiqiang Li, and Wei Fan, Mingyu Yang, Jinsen Li, Xuesong Hu, Ping Zhu, Dr. Liya Xu and Prof. Fan Bai for their contributions to the project mentioned in Chapter 6.

I thank our volleyball team RamAm Noodles for the great games in the past five years. Playing with RamAm Noodles has been more than just having fun. We developed the mental toughness and built up mutual trust in the games, which have helped me greatly.

I thank Professors Xiaowei Zhuang and Adam Cohen for serving on my dissertation committee. I am grateful to Professor Cohen for a chance to rotate in his lab in the winter of 2007, and I recalled lying on the floor at 2am discussing chirality, with our arms and legs wrapped around a meter-long stick for showing the electric and magnetic fields. I truly treasure the excitement and the special experience that I could have never gotten in any other place in the world.

I am grateful to my parents, who have been supporting me unconditionally throughout my life. I thank Ms. Jiejun Chen, for accompanying me since the last century, and for helping me through the difficult moments. I could have never done this without their support.

Part I (Chapters 1-4):

Label-Free Optical Imaging of Chromophores

Introduction

Optical imaging is an essential tool for biomedical research because of its unique properties — high resolution, in-vivo detection capability and versatile contrast mechanisms. Fluorescence Microscopy has been widely used to explore the nanoscale cellular world because of its superb sensitivity and the availability of modern labeling technologies. However, many species (e.g. lipids, nucleic acids, small proteins) are difficult to label without affecting their native functions. To study the intracellular dynamics of these species, label-free imaging techniques need to be developed with high sensitivity and resolution. Most of these species have undetectable fluorescence under visible illumination. Therefore new contrast mechanisms need to be explored for label-free imaging of these species.

In this part of the thesis, we explore three different contrast mechanisms for label-free imaging of biological materials — absorption and stimulated emission (Chapter 2: Stimulated Emission Microscopy), heat generation and diffusion (Chapter 3: Two-Photon Excited Photothermal Lens Microscopy), and nonlinear scattering (Chapter 2: Near-Degenerate Four-Wave-Mixing Microscopy). We demonstrate label-free imaging of blood vessels, cytochromes, drugs for photodynamic therapy, muscle and brain tissues with three dimension optical sectioning capability.

Contributions

This part of the thesis involved close collaboration with Dr. Wei Min.

In Chapter 2: Stimulated Emission Microscopy, Dr. Min, I and Prof. Xie conceived the idea and designed the experiments. Dr. Min, I and Shasha Chong performed experiments and analyzed data. Dr. Rahul Roy constructed the E. Coli cells expressing chromoproteins, Dr. Holtom and S. Chong helped to construct the laser systems.

In Chapter 3: Two-photon Photothermal Lens Microscopy, Dr. Min, I and Prof. Xie conceived the idea and designed the experiments. Dr. Min, I and S. Chong performed the experiments and analyzed data. Dr. Holtom helped with the laser source.

In Chapter 4: Near-Degenerate Four-Wave-Mixing Microscopy, Dr. Min, I and Prof. X. Sunney Xie conceived the idea and designed the experiments. Dr. Min and I performed the experiments and analyzed data. Dr. Markus Rueckel and Dr. Gary R. Holtom helped with the laser source and the microscope setup.

Chapter 1

Introduction to Label-Free Optical Imaging

1.1 The Interaction of Light and Matter

The interaction of light and matter is one of the most fundamental forms of interaction in nature. Light as particles interacts with matter by transferring quantized energy and momentum, and light as waves interacts through electric and magnetic field driving the oscillation of the matter (McHale, 1998). In the presence of an external field, the response of the matter (represented by the dipole moment) can be expanded in a power series of the electric field:

$$\vec{\mu} = \vec{\mu_0} + \alpha \cdot \vec{E} + \frac{1}{2}\beta : \vec{E}\vec{E} + ... \qquad (1.1)$$

The constant term $\overrightarrow{\mu_0}$ is the permanent dipole moment, which does not respond to the field. The linear term introduces linear polarizability α, which is related to the linear scattering and absorption of light. The nonlinear terms represent higher order polarizability of the matter, which are often negligible under normal light intensity but can be significant with high power pulse laser.

Looking into the molecular detail of the matter, the polarizability of the system can be described by a collection of N elections, each with a harmonic frequency ω_j.

$$\alpha = \frac{e^2}{m}\sum_j \frac{f_i}{\omega_j^2} \qquad (1.2)$$

The quantity f_i is the oscillator strength. In the case where the filed is a function of time and has a frequency of ω, after solving the equation of a driven harmonic oscillator, the polarizability of the system can be described as:

$$\alpha(\omega) = \frac{e^2}{m}\sum_j \frac{f_i}{\omega_j^2 - \omega^2 - i\omega/\tau} \qquad (1.3)$$

The quantity τ represents the relaxation time for damping of the induced dipole, which correspond to the state lifetime in quantum mechanical description of molecular states. The oscillator strength f_i is related to the transition dipole of the states:

$$\alpha(\omega) = \frac{e^2}{m}\sum_j \frac{f_{0j}}{\omega_{0j}^2 - \omega^2 - i\omega/t} \quad \text{with} \quad f_{0j} = \frac{2m\omega_{0j}\mu_{0j}^2}{3e^2\hbar} \tag{1.4}$$

If we have the information of these molecular details, i.e., the quantities described above, in principle the linear scattering $\mathrm{Re}[\alpha(\omega)]$ and the absorption $\mathrm{Im}[\alpha(\omega)]$ can be calculated. They are related by the Kramers-Kronig relations. Vice versa, by measuring the linear scattering or absorption, molecular properties such as the energy states and transition dipole strength can be inferred. This sets up the foundation for optical microscopies based on linear scattering and absorption (Pawley, 2006).

1.2 Fluorescence Imaging. Why and Why Not Enough?

Although linear scattering and absorption have been widely used to study molecular spectroscopy (McHale, 1998), the usage of these contrast mechanisms for ultra-sensitive optical imaging has been hampered by several technical difficulties. First, the optical signal exists as a modulation of the input optical field, and the signal needs to be detected in the background of the input light. As a rough estimation, the absorption cross section of a molecule is on the order of (0.1nm)^2, and the optical resolution limit is about (300nm)^2. When a tightly focused optical field passes through a molecule sitting at the focus of the field, only about (0.1nm)^2/(300nm)^2~ 10^-7 of the photons are absorbed by the molecule, which is normally undetectable in the background of the incidence optical field. Second, absorption and linear scattering do not naturally offer 3-D optical sectioning, and the detected optical

signal reflects a cumulative effect along the input optical field unless using complex interference methods (Huang et al., 1991). Therefore, the images are often blurred when dealing with thick biological specimen.

Fluorescence offers a natural solution to these technical issues (Lakowicz, 2006). As is shown in Figure 1.1, fluorescence photons exhibit lower energy compared to the input excitation photon. By using an optical filter to selectively detect the fluorescence wavelength, fluorescence detection can be done essentially background-free. Fluorescence detection was demonstrated to have single molecule sensitivity in room temperature (Trautman et al., 1994; Xie and Dunn, 1994), which has been widely used to study molecular dynamics in vitro (Selvin and Ha, 2007) and in vivo (Li and Xie, 2011).

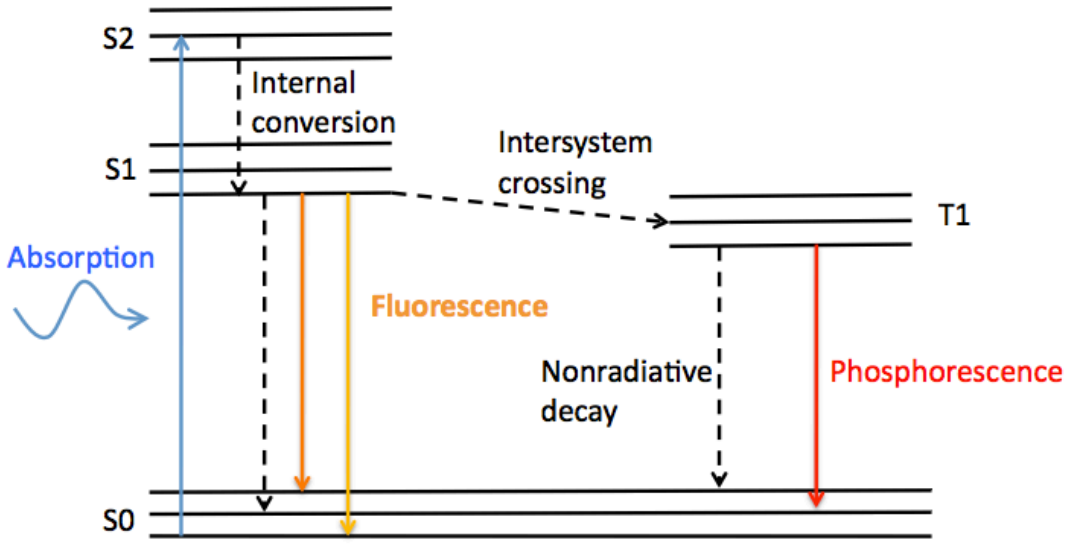


Figure 1.1 Jabcoski diagram showing the processes of absorption and fluorescence

The relatively long excited state lifetime (~ns) of the fluorescence molecules was used to saturate the mission that allowed breaking the resolution limit in optical microscopy (Hell and Wichmann, 1994). The single molecule detection sensitivity was later used to significantly simplify the instrumentation for super resolution imaging (Betzig et al., 2006; Rust et al., 2006) which is currently widely used to study biological dynamical structure that were no resolvable before (Huang et al., 2010). The background-free nature of the fluorescence detection also allows 3-D optical sectioning by the development of confocal fluorescence microscopy (2006) and two-photon fluorescence microscopy (Denk et al., 1990), which is essential in imaging through thick biological specimen.

Another major reason for the wide usage of fluorescence in biological research is the versatile pool of labeling techniques developed for different species. The developments of fluorescence proteins allow genetically encodable labeling with high specificity in living cells and animals (Chalfie et al., 1994; Tsien, 1998). The recent developments of techniques for labeling small protein and other molecules such as RNA is also quickly expanding the pool of available labeling tools (Paige et al., 2012).

Despite the wide usage in biological research, fluorescence has limitations that necessities the development other optical techniques. First, most molecular species are intrinsically non-fluorescent or have very weak fluorescence. Fluorescence labeling is often very challenging for species such as small peptides, metabolites, lipids, and drugs. It is also difficult to guarantee the labeling does not affect the normal function of the molecules, especially when the size of the label exceeds that of the molecules under study. Moreover, for medical applications especially in human studies, it is preferable no external labels are applied because of the toxicity and side effects of fluorescence dyes.

1.3 Contrast Mechanisms and Label-Free Optical Imaging

The various forms of light-matter interactions offer a variety of contrast mechanisms for label-free optical imaging (Min et al., 2011). The linear polarizability α gives rise to linear scattering and absorption, which offers the contrast mechanism for Optical Coherence Tomography (Huang et al., 1991), a method now widely used for in vivo medical imaging of the 3-D structure of various biological tissues such as retina.

With the developments of ultrafast high power laser systems, higher order terms of equation 1.1 were explored to offer different contrast mechanisms for label-free chemical imaging (Boyd, 2008). The second order polarizability gives rise to second harmonic generation (SHG), which was used to image highly ordered structures such as collagen (Campagnola et

al., 2002). SHG is surface sensitive and was also used to image the electric activities on cellular membranes (Dombeck et al., 2004).

The third order polarizability gives rise to phenomena such as third harmonic generation, coherent Raman scattering and two-photon absorption (Boyd, 2008). Coherent anti-stokes Raman scattering (CARS) microscopy was developed to have molecular selectivity based on its Raman spectrum (Zumbusch et al., 1999), and has been widely used to study lipids and other substances (Cheng and Xie, 2004; Evans and Xie, 2008). The sensitivity of coherent Raman scattering microscopy is further improved by measuring stimulated Raman scattering (SRS), and was demonstrated to be able to distinguish different types of lipids (Freudiger et al., 2008). SRS has been demonstrated to have video rate scanning speed in thick human tissues (Saar et al., 2010), which allows in vivo chemical imaging in living animals and humans.

1.4 An Overview of Chapters 2-4

In Chapter Two, we explore the usage of linear polarizability as a contrast mechanism for sensitive high resolution imaging. As discussed in 1.2, absorption and linear scattering detection are not background-free in microscopy and often have low sensitivity and resolution. Therefore, instead of measuring absorption directly, we measure the reverse process stimulated emission, and we use the modulation transfer method to suppress the fluctuation of the background light. The achieved high sensitivity and 3-D sectioning

capability allow us to demonstrate imaging non-fluorescent proteins, drug distribution and microvascular structures (Min et al., 2009a).

In Chapter Three, we continue the study of using absorption for label free imaging. Here we explore using the heat effect of the absorption as a contrast mechanism for imaging heme derivatives such as cytochromes and hemoglobin. Unlike in chapter two, we do not need synchronized ultrafast laser systems and therefore the technique can be used as a general detection scheme for absorption detection in microscopy (Lu et al., 2010).

In Chapter Four, we explore using a new contrast mechanism, near degenerate four-wave mixing, for label free imaging of the molecules that have low absorption coefficient in the visible wavelengths. We demonstrate imaging thick biological samples such as muscles and brain tissues (Min et al., 2009b).

References:

Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., and Hess, H.F. (2006). Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. Science *313*, 1642–1645.

Boyd, R.W. (2008). Nonlinear Optics, Third Edition (Academic Press).

Campagnola, P.J., Millard, A.C., Terasaki, M., Hoppe, P.E., Malone, C.J., and Mohler, W.A. (2002). Three-Dimensional High-Resolution Second-Harmonic Generation Imaging of Endogenous Structural Proteins in Biological Tissues. Biophysical Journal *82*, 493–508.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., and Prasher, D.C. (1994). Green fluorescent protein as a marker for gene expression. Science *263*, 802–805.

Denk, W., Strickler, J.H., and Webb, W.W. (1990). Two-photon laser scanning fluorescence microscopy. Science *248*, 73–76.

Dombeck, D.A., Blanchard-Desce, M., and Webb, W.W. (2004). Optical Recording of Action Potentials with Second-Harmonic Generation Microscopy. J. Neurosci. *24*, 999–1003.

Evans, C.L., and Xie, X.S. (2008). Coherent anti-Stokes Raman scattering microscopy: chemical imaging for biology and medicine. Annu. Rev. Anal. Chem. *1*, 883–909.

Freudiger, C.W., Min, W., Saar, B.G., Lu, S., Holtom, G.R., He, C., Tsai, J.C., Kang, J.X., and Xie, X.S. (2008). Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy. Science *322*, 1857–1861.

Giepmans, B.N.G., Adams, S.R., Ellisman, M.H., and Tsien, R.Y. (2006). The Fluorescent Toolbox for Assessing Protein Location and Function. Science *312*, 217–224.

Grynkiewicz, G., Poenie, M., and Tsien, R.Y. (1985). A new generation of Ca2+ indicators

with greatly improved fluorescence properties. Journal of Biological Chemistry *260*, 3440–3450.

Hell, S.W., and Wichmann, J. (1994). Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Optics Letters *19*, 780–782.

Huang, B., Babcock, H., and Zhuang, X. (2010). Breaking the Diffraction Barrier: Super-Resolution Imaging of Cells. Cell *143*, 1047–1058.

Huang, D., Swanson, E.A., Lin, C.P., Schuman, J.S., Stinson, W.G., Chang, W., Hee, M.R., Flotte, T., Gregory, K., Puliafito, C.A., et al. (1991). Optical coherence tomography. Science *254*, 1178–1181.

Lakowicz, J.R. (2006). Principles of Fluorescence Spectroscopy (Springer).

Li, G.-W., and Xie, X.S. (2011). Central dogma at the single-molecule level in living cells. Nature *475*, 308–315.

Lu, S., Min, W., Chong, S., Holtom, G.R., and Xie, X.S. (2010). Label-free imaging of heme proteins with two-photon excited photothermal lens microscopy. Applied Physics Letters *96*, 113701–113701–3.

McHale, J.L. (1998). Molecular Spectroscopy (Prentice Hall).

Min, W., Freudiger, C.W., Lu, S., and Xie, X.S. (2011). Coherent Nonlinear Optical Imaging: Beyond Fluorescence Microscopy. Annual Review of Physical Chemistry *62*, 507–530.

Min, W., Lu, S., Chong, S., Roy, R., Holtom, G.R., and Xie, X.S. (2009). Imaging chromophores with undetectable fluorescence by stimulated emission microscopy. Nature *461*, 1105–1109.

Paige, J.S., Wu, K.Y., and Jaffrey, S.R. (2011). RNA Mimics of Green Fluorescent Protein. Science *333*, 642–646.

Pawley, J. (2006) Handbook of Biological Confocal Microscopy (Springer).

Rust, M.J., Bates, M., and Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nature Methods *3*, 793–796.

Saar, B.G., Freudiger, C.W., Reichman, J., Stanley, C.M., Holtom, G.R., and Xie, X.S. (2010). Video-Rate Molecular Imaging in Vivo with Stimulated Raman Scattering. Science *330*, 1368–1370.

Selvin, P.R. and Ha, T (2007) Single-Molecule Techniques: A Laboratory Manual (Cold Spring Harbor Laboratory Press).

Trautman, J.K., Macklin, J.J., Brus, L.E., and Betzig, E. (1994). Near-field spectroscopy of single molecules at room temperature. , Published Online: 05 May 1994; | Doi:10.1038/369040a0 *369*, 40–42.

Tsien, R.Y. (1998). The green fluorescent protein. Annual Review of Biochemistry *67*, 509–544.

Xie, X.S., and Dunn, R.C. (1994). Probing Single Molecule Dynamics. Science *265*, 361–364.

Zumbusch, A., Holtom, G.R., and Xie, X.S. (1999). Three-Dimensional Vibrational Imaging by Coherent Anti-Stokes Raman Scattering. Phys. Rev. Lett. *82*, 4142–4145.

Chapter 2

Stimulated Emission Microscopy

2.1 Introduction

Fluorescence has been generally considered to be more sensitive than absorption measurement in optical imaging (Pawley, 2006). However, many important biological chromophores, such as cytochromes and hemoglobin, absorb light but have undetectable fluorescence, because the spontaneous emission is dominated by their fast non-radiative decay rates (Turro, 1991). However, conventional one-beam absorption measurement exhibits low sensitivity, lack of three-dimensional sectioning capability, and complication by linear scattering of heterogeneous samples. Here we use stimulated emission, which competes effectively with the nonradiative decay to make these chromophores detectable, and we report a new contrast mechanism for optical microscopy.

In a pump-probe experiment, upon photoexcitation by a pump pulse, the molecules is stimulated down to the ground state by a time-delayed probe pulse, the intensity of which is concurrently increased. We modulate the intensity of the pump beam at a high megahertz frequency, and we extract the miniscule intensity increase of the probe beam with shot-noise-limited by using a lock-in amplifier. The signal is generated only at the laser focus owing to the nonlinear dependence on the input intensities, which provides intrinsic three-dimensional optical sectioning capability. We demonstrate various applications of stimulated emission imaging, such as visualizing chromoproteins, non-fluorescent variants of the green fluorescent protein, monitoring lacZ gene expression with a chromogenic reporter, mapping transdermal drug distributions without histological sectioning, and label-free microvascular imaging based on endogenous contrast of hemoglobin. For all these applications, sensitivity is orders of magnitude higher than for spontaneous emission or absorption contrast, permitting nonfluorescent reporters for molecular imaging. The descriptions in this chapter is based on a previously published work (Min et al., 2009)

2.2 Stimulated Emission and Estimation of Signal Strength

The phenomenon of stimulated emission was first described by Einstein in 1917 (Einstein, 1917). A molecule in its excited state can be stimulated down to the ground state by an incident light field if the energy of the light field matches with the energy difference between

molecular states, resulting in the creation of a new coherent photon identical to those in the incident field. Stimulated emission was later used as a fundamental principle for light amplification in the laser (Seigman, 1986). The depopulation aspect of stimulated emission has been used for population dumping from excited states (Hamilton et al., 1986), super-resolution fluorescence microscopy (Hell and Wichmann, 1994), and fluorescence lifetime imaging (Dong et al., 1995). Here we use the light-amplification aspect of stimulated emission as a contrast mechanism for sensitive imaging of chromophores that have undetectable fluorescence. These chromophores have very short-lived excited states with much faster non-radiative decay rates than their spontaneous emission rates. As a result, their feeble fluorescence is overwhelmed by backgrounds such as stray light, solvent Raman scattering, and detector dark counts.

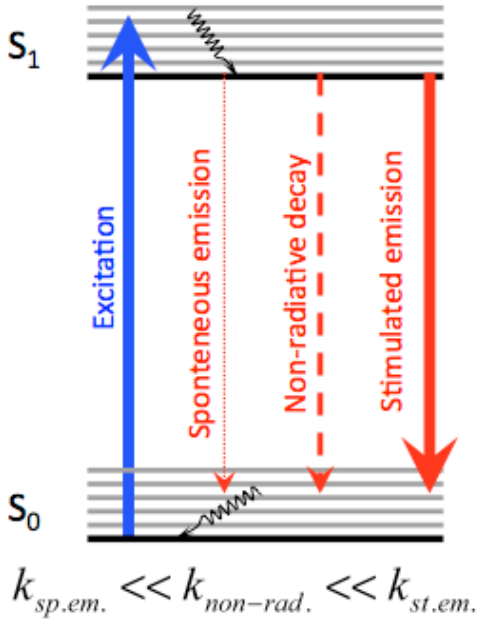


Figure 2.1: Principle of using stimulated emission to compete with the fast non-radiative decay rate of the chromophores. For these molecules, spontaneous emission is very weak because of the domination of non-radiative decay rate. However if the stimulated emission field is strong enough during the lifetime of S1, the rate of stimulated emission can be much faster than non-radiative decay rate.

Our solution to this problem is to conduct a dual-beam system to probe the short-lived excited state by stimulated emission, which can compete with the non-radiative decay under a strong enough stimulating field (Figure 2.2). The resulting 'amplification' of the stimulation beam can then be detected in the presence of the background signals.

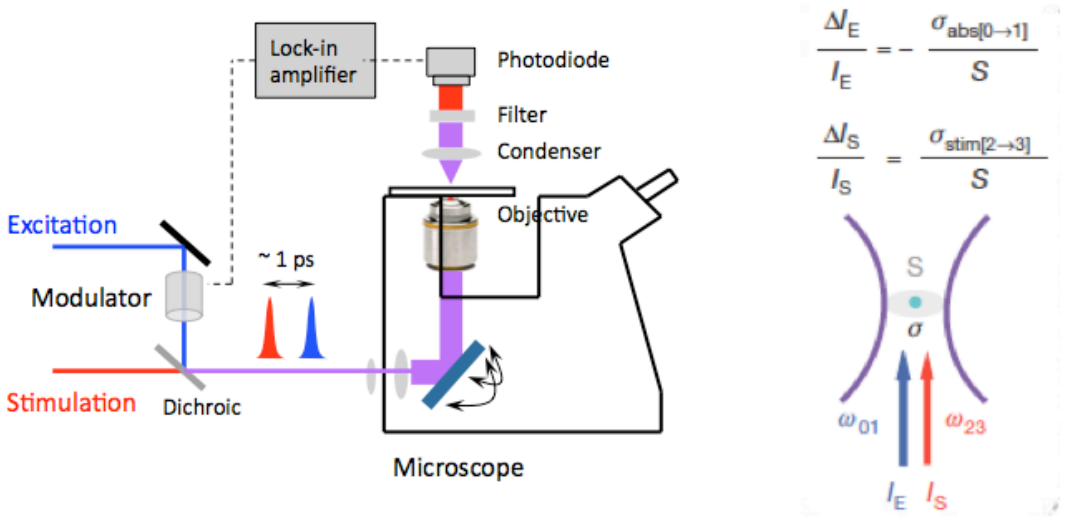


Figure 2.2: A simple scheme of the stimulated emission microscope. We used two beams to excite and to induce stimulated emission of the molecule.

Considering the optical excitation at frequency ω_{01} (Figure 2.2), the absorption cross section $\sigma_{abs[0\to1]}$ is ~ 10^{-16} cm^2 for a single chromophore at room temperature (Cantor and Schimmel, 1980; Lakowicz, 2006). As shown in Figure 2.2, under a tightly focused laser beam with a waist area of S (~ 10^{-9} cm^2 for visible light focused by a high N.A. objective), the integrated intensity attenuation of the excitation beam, $\Delta I_E / I_E$, is proportional to the ratio between $\sigma_{abs[0\to1]}$ and S:

$$\Delta I_E / I_E = -N_0 \cdot \sigma_{abs[0\to1]} / S \qquad (2.1)$$

where N_0 is the number of molecules in the ground state. $\Delta I_E / I_E$ is on the order of 10^{-7} for a chromophore with large absorption cross section. Such small attenuation cannot be detected by conventional absorption microscopy. We note that single molecule absorption was previously achieved in cryogenic temperatures using a frequency modulation method (Moerner and Kador, 1989) which is however difficult to implement at room temperature because of the broad molecular absorption spectrum. Instead of detecting the direct absorption, here we detect the stimulated emission followed by the excitation of the molecule.

According to Einstein (Einstein, 1917), the molecular cross section $\sigma_{sti.em.}$ for stimulated emission is comparable to σ_{abs}, because of microscopic reversibility. Unlike the absorption that results in intensity attenuation, the stimulation beam experiences an intensity gain after interacting with the molecules

$$\Delta I_S / I_S = +N_2 \cdot \sigma_{sti.em.[2\to 3]} / S \qquad (2.2)$$

N_2 is the number of excited molecules probed by the stimulation pulses. For a single chromophore residing in the excited state, $\Delta I_S / I_S$ is also ~10^{-7}. Without special techniques, such a small signal would be again buried in the noise (~1%) of the stimulation beam.

2.3 Detection Scheme and Experimental Setup

To overcome this noise problem for detecting stimulated emission from a small number of molecules in microscopy, we implemented a high-frequency (> 1 MHz) phase-sensitive detection technique, as shown in Figure 2.3.

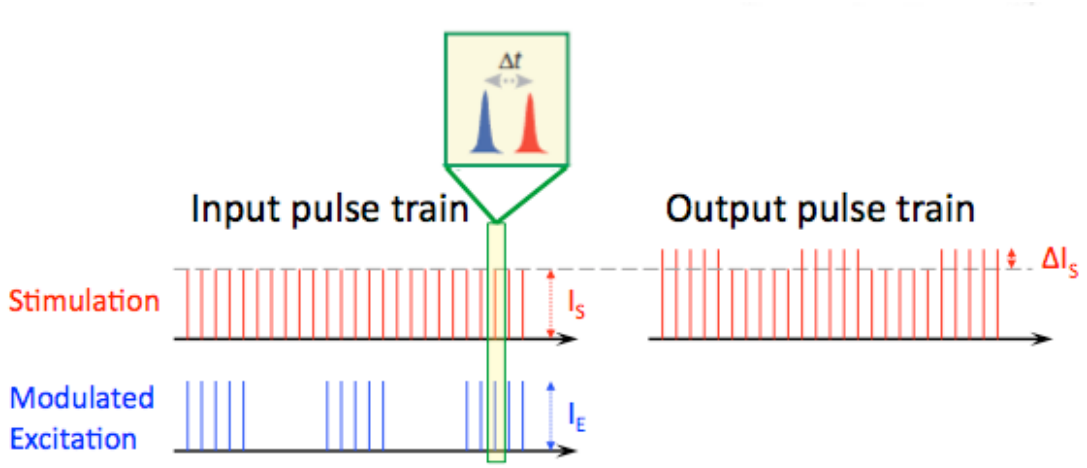


Figure 2.3: The detection scheme of a stimulated emission microscope. As the intensity of the excitation beams is modulated on and off over time, the corresponding stimulated emission signal is also modulated at the same frequency, which can be demodulated with a lock-in amplifier. The stimulated pulses are made to lag behind the excitation pulses by a $\Delta t \sim 1$ps.

In so doing, the laser intensity fluctuation, which occurs primarily at low frequencies (kHz to DC), can be circumvented, as has been previously applied in other spectroscopic (Ye et al., 1998) and microscopic applications (Freudiger et al., 2008). The intensity of the excitation beam is modulated at 5 MHz, and this creates a modulation of the stimulated emission signal at the same frequency, because only when the excitation beam is present can the gain of the stimulation beam occur. Such an induced modulation signal can then be extracted by a lock-in

amplifier referenced to this high frequency. In this way, the dual beam modulation transfer scheme detects the stimulated emission signal against the vanishing laser noise at a high frequency, offering a superior sensitivity over the direct one-beam absorption detection scheme. We use a ~200 fs (FWHM) pulse train for excitation, and another ~200 fs pulse train for stimulation. The time delay between these two pulse trains is chosen to be ~300 fs, which is shorter than the excited state lifetime (sub-ps) of the chromophores. This delay also eliminates contributions from other instantaneous optical processes, such as two photon absorption (Fu, 2007), cross phase modulation and stimulated Raman scattering (Freudiger et al., 2008).

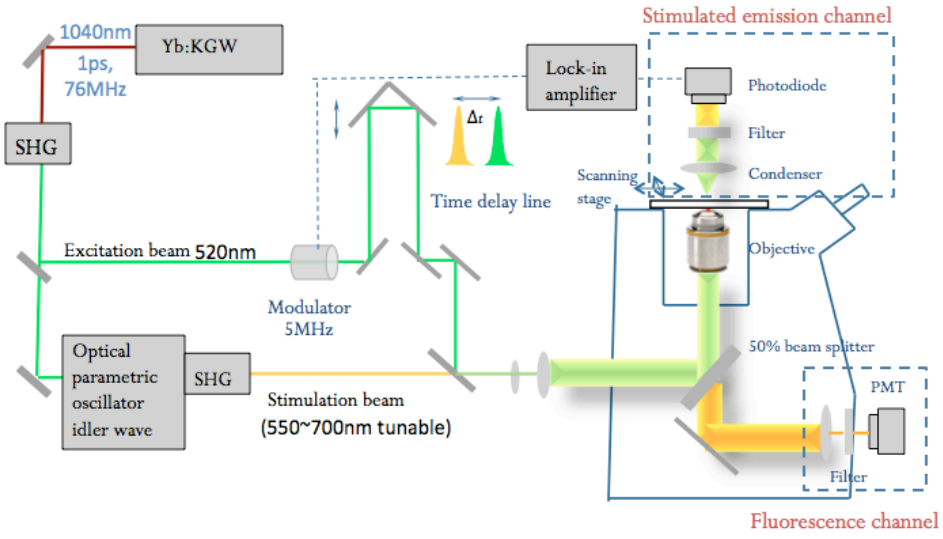


Figure 2.4: The detailed experimental setup of a stimulated emission microscope. Another scheme of operation is to use a Ti-sapphire oscillator to pump two Optical parametric oscillators, which was used in the early phase of this project.

Specifically, a 10W solid-state 532nm green laser (Millennia, Spectral-Physics) is used to pump a Ti-sapphire oscillator (Mira, Coherent) to produce a 2W mode-locked 100fs 76MHz pulse train at 830nm. This pulse train is then split and used to synchronously pump two fs optical parametric oscillators (OPOs) simultaneously: the first OPO's (Mira-OPO basic, Coherent) output wavelength is tuned by tuning the color of the pump Ti-sapphire output (under 830nm Ti-sapphire pumping, this OPO signal wave is at 1180nm), and the second OPO (Mira PP-OPO, Coherent) is tuned by adjusting the cavity length of the OPO (the range of the signal wave is between 1100nm and 1400nm). The two independent frequency-doubled outputs from these two OPO signal waves, in the wavelength range of 560 to 700 nm with pulse widths around 200 fs, serve as the either excitation or stimulation pulse trains. A pulse compressor consisting of a pair of SF11 prisms is built to control the pulse width. Collinear excitation and stimulation beams are combined and focused with a high numerical aperture (NA=1.2) objective onto a common focal spot. The temporal delay between the synchronized excitation and stimulation inter-pulse is adjusted to between 0.2 and 0.3 ps. The intensity of the excitation beam is modulated by an acousto-optical modulator (Crystal technology) at 5 MHz. A condenser with a N.A.=0.9 is used to collect the forward propagating stimulation beam, which is spectrally filtered before detected by a photodiode. To acquire images with a laser scanning microscope (FV300, Olympus), we used a 100 μs time constant for a lock-in amplifier (SR844, Stanford Research) and pixel dwell time of 190 μs.

For imaging of chromoproteins, X-gal hydrolysis product and TBO drug distributions, the first OPO output is frequency doubled outside the cavity by a BBO crystal to generate a fs pulse train around 590nm as the excitation beam, and the second OPO is intra-cavity doubled to generate its second harmonic signal between 550nm and 700nm as the stimulation beam. For imaging blood vessels, 830nm Ti-sapphire serves as the excitation beam and the frequency-doubled Mira PP-OPO around 600nm served as the stimulation beam.

The excitation beam and the stimulation beam are spatially overlapped with a dichroic beam splitter. Temporal delay between two excitation and stimulation pulse trains is set with a translation delay-stage and measured with an autocorrelator (APE GmbH). The exact time zero is adjusted by optimizing the coherent anti-Stokes Raman scattering signal around 534nm generated by the pump beam at 590nm and Stokes beam at 660nm.

The excitation beam is modulated by an acousto-optical modulator (AOM) (model 3080-122, Crystal technology) at 5 MHz which is driven by a square-wave function generator. We note that the AOM crystal adds significant chirp to the pulses. To compensate for this, a pulse compressor consisting of a pair of SF11 prisms (Thorlabs) is built into the excitation beam path to control its pulse width.

Excitation and stimulation beams are coupled into a modified laser scanning inverted microscope (IX71, FV300, Olympus). The beam size is matched to fill the back-aperture of objective. A 60X 1.2 N.A. water objective (UPlanSApo, Olympus) is used for excitation, and a 20X 0.95 N.A. long-working distance objective (XLUMPlanFI, water, Olympus) is used as a condenser. Another lens is used to image the scanning mirrors onto a silicon amplified photo-diode (PDA36A, Thorlabs) to avoid beam movement during laser scanning. Two high OD bandpass filters (HG650/45X, Chroma Technology; Brightline fluorescence filter 655/40, Semrock) are used together to block the excitation beam completely and only transmit the stimulation beam. For imaging blood vessels, high OD filters (3^{RD}800SP and 3^{RD} 760SP, Omega Optical) are used together to block the excitation beam completely.

The output of the photodiode is bandpass filtered (15542, DC-48MHz low-pass filter, Mini-Circuits) to suppress the strong signal at the pulsing repetition rate (76 MHz), and then terminated with 50Ω. A high-frequency lock-in amplifier (SR844, Stanford Research) is used to demodulate the stimulated emission signal. The analog on phase component x-output of the lock-in amplifier is fed into the A/D converter of the microscope input. The time constant is set for 1 sec and 100 μs under spectroscopy and microscopy experiments, respectively.

We conduct the experiment under a non-saturating condition of the four-level system shown in Figure 2.1. Under this condition, N_2 in equation 2.1 originates from a linear excitation: $N_2 \propto N_0 \cdot I_E \cdot \sigma_{abs[0\to1]}/S$. This relation, together with equation 2-2, indicates that the final signal ΔI_S is linearly dependent on both I_E and I_S,

$$\Delta I_S \propto N_0 \cdot I_E \cdot I_S \cdot (\sigma_{abs[0\to1]}/S) \cdot (\sigma_{sti.em.[2\to3]}/S) \tag{2.3}$$

The overall quadratic power dependence, as is experimentally demonstrated (Figure 2.5), would allow three-dimensional (3D) optical sectioning, as in many other multi-photon techniques (Denk et al., 1990; Evans and Xie, 2008). Moreover, it offers, in principle, a spatial resolution of twice as high (in spatial frequency) as in conventional fluorescence microscopy.

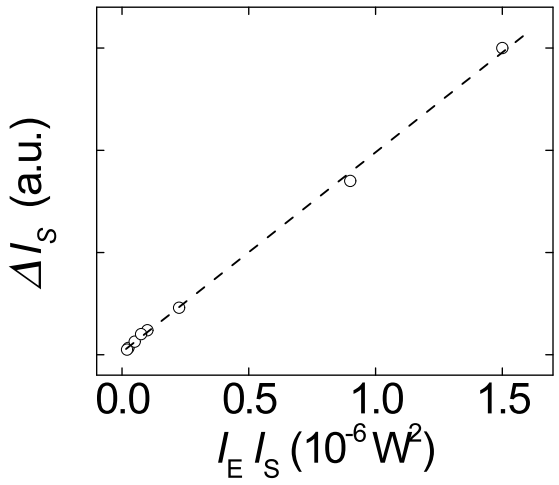


Figure 2.5 Linear dependence of stimulated emission signal, ΔI_S, on the product of excitation beam power, I_E, and stimulation beam power, I_S.

2.4 Signal Characterization

We first look at the stimulated emission signal as a function of the time-delay between the excitation and stimulation pulses. The initial rise is due to vibrational relaxation shown in Figure 2.1, while the subsequent decay indicates the short excited state lifetime (~0.6 ps) of the molecule, which underlies the non-detectable fluorescence. Such a short lifetime also reduces the probability of going into the triplet state, effectively protecting the molecule from photo-bleaching.

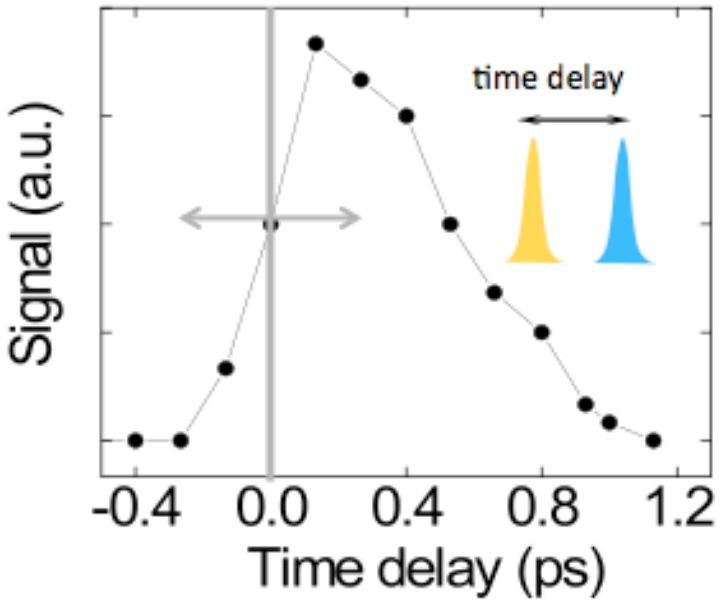


Figure 2.6 Time response of the stimulated emission signal measured by tuning the time delay between the pump and the probe beam

We then look at the spectrum of the stimulated emission signal (Figure 2.7), recorded by tuning the wavelength of the stimulated beam.

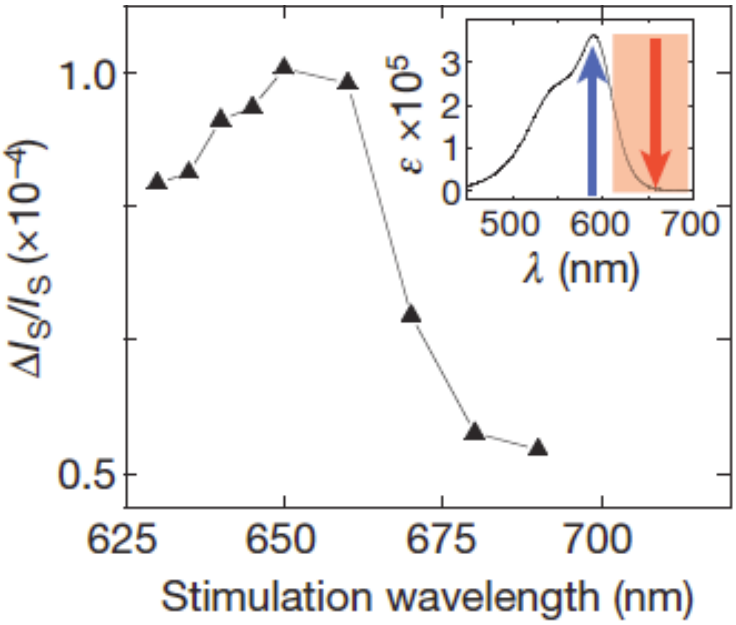


Figure 2.7. The spectrum of the stimulated emission signal measured by tuning the wavelength of the stimulation beam. The 590nm excitation is fixed while the stimulation wavelength is canned by tuning the other optical parametric oscillator wavelength. A time delay of 0.3ps is used.

The wavelength dependence is consistent with the reported emission spectrum of crystal violet in glycerol (Du, 1998) in which the high viscosity increases the fluorescence quantum yield.

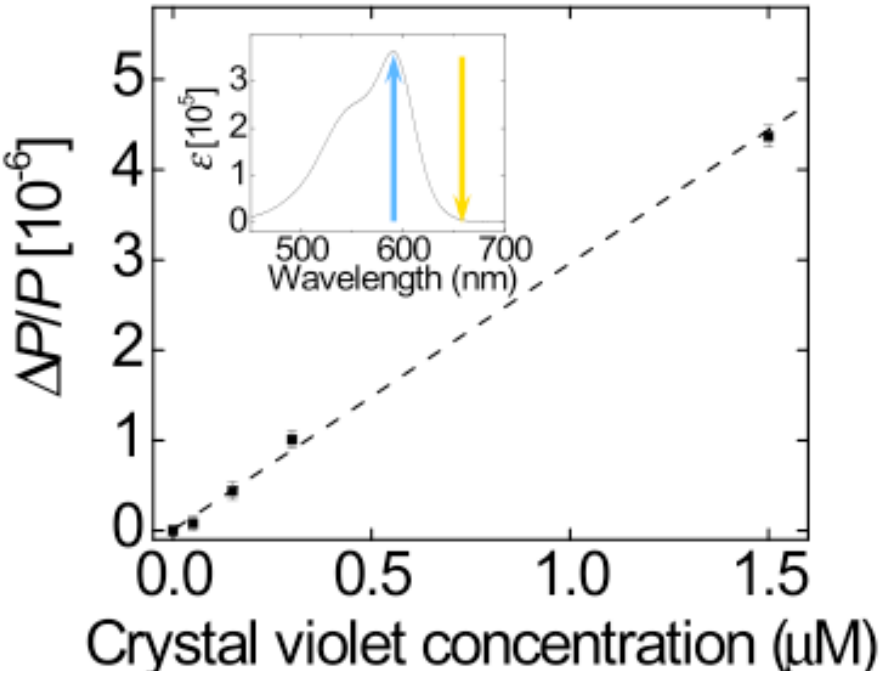


Figure 2.8. The stimulated emission signal scales linearly with the crystal violet concentration in aqueous solution. Continuous flow of the sample is used to replenish molecules. Wavelengths are 590nm and 660nm for excitation and stimulation beam, respectively, and the time delay of 0.3ps is used. Error bar show 1 s. d. of the signals from a 30s recording. The detection limit was determined to be 60nM with a signal-to-noise ratio of 1:1.

Figure 2.8 shows that the stimulated emission signal depends linearly on analyte concentration, as predicted by equation (2.3). This allows straightforward quantitative analysis. The limit of detection ($\Delta I_S / I_S \sim 10^{-7}$) is ~60 nM for crystal violet with 1 sec integration time. Approaching the shot noise limit, this sensitivity corresponds to a few (<5)

molecules in focus, which has surpassed the detection limit of recently reported for resonant nonlinear scattering microscopy by two orders of magnitude (Du, 1998).

High purity crystal violet powder is used as purchased (Sigma Aldrich). Aqueous solutions are prepared with de-ionized water. For spectroscopy, we built a flow-cell from two No.1 coverslips and a spacer to allow quick concentration exchange without moving the sample position or focusing depth inside the sample. The absolute concentration is checked by a UV-vis spectrophotometer.

2.5 Three Dimensional Optical Sectioning

The general intensity dependence of stimulated emission is quadratic; therefore, we expect a steep decrease of signal moving away from the focus, which gives 3-D optical sectioning capability. Here we demonstrate the 3-D sectioning using a hematoxylin and eosin (H&E) stained slice of gray matter from a mouse brain. We took a one-micron step for each z-section to construct the 3-D image.

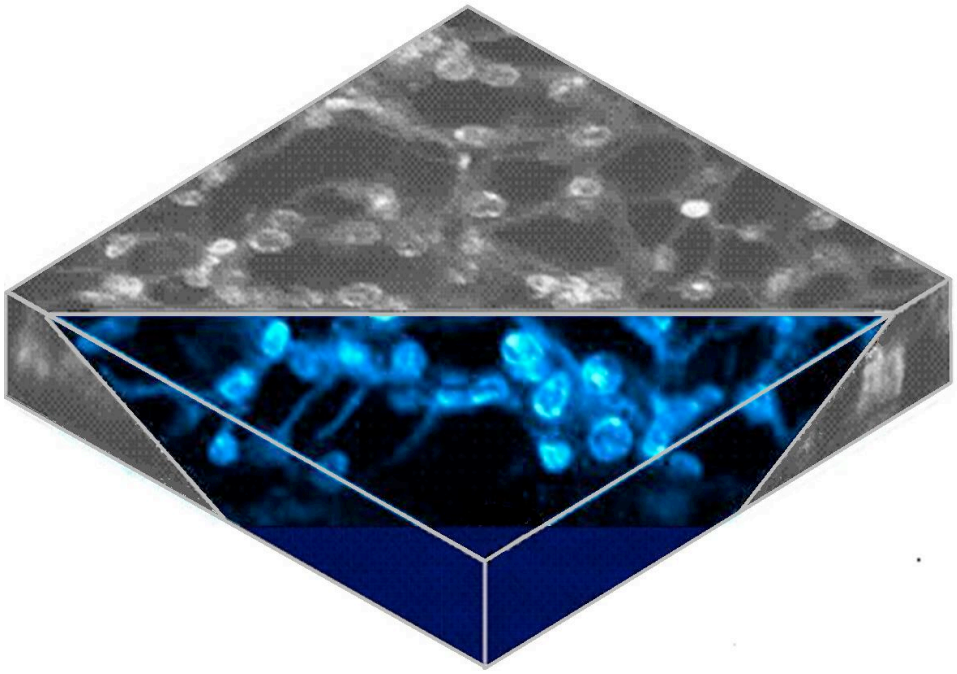


Figure 2.9 Three dimensional sectioning capability of stimulated emission microscope. Here we show a z-stack constructed 3-D image from the gray matter of a mouse brain stained by Hematoxylin and eosin.

2.6 Imaging Non-Fluorescent Chromoproteins and Chromogenic Reporter

As the first biological application, we image distributions of chromoproteins in live *E. coli* cells. Genetically encodable chromoproteins, such as gtCP (Gurskaya, 2001) and cjBlue (Chan, 2006) , are variants of green fluorescent proteins (Zhang et al., 2002), and only absorb light but do not fluoresce. When the gene encoding for gtCP is expressed in *E. coli* cells, tetrameric gtCP can be clearly visualized to reside homogeneously inside cytoplasm by

stimulated emission microscopy. Similarly, the distribution of another chromoprotein, cjBlue, can be imaged. Stimulated emission microscopy opens the possibility of utilizing chromoproteins as genetically encodable imaging probes.

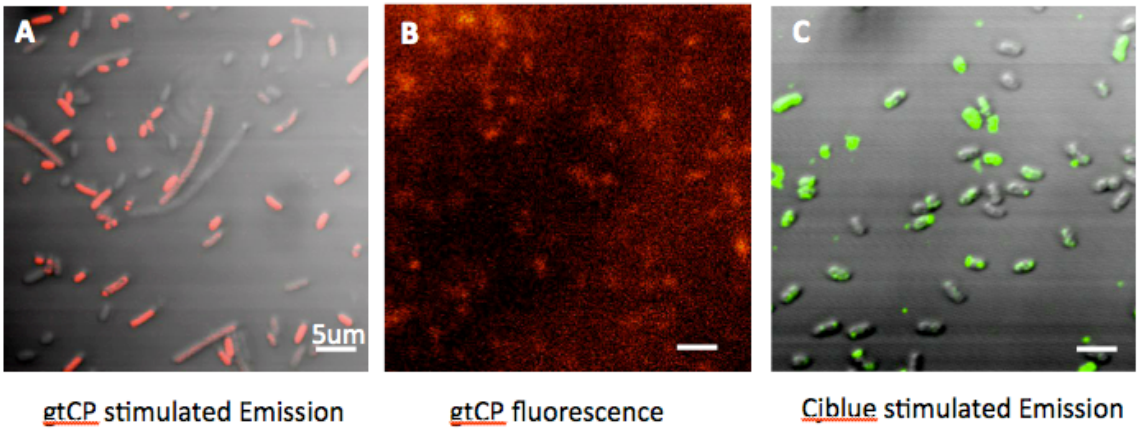


Figure 2.10 Imaging non-fluorescent chromoproteins by stimulated emission. Imaging distributions of cytoplasmic chromoproteins gtCP (a) and cjBlue (c) in live E. coli cells, respectively, by stimulated emission microscopy, overlapped with the corresponding wide-field transmission images. (b) shows the fluorescence yield for gtCP is low, therefore the image quality is low due to stray light and high gain of the PMT.

Chromoprotein gtCP was expressed in the DH10B *E. coli* using pQE30 expression vector without induction. After cell growth in LB medium at 37 °C to A_{600} of 0.6, the culture was moved to 22°C shaker for 24 -36 hrs to ensure complete maturation of the chromophore. cjBlue was overexpressed from a pRSETB vector in BL21(DE3) *E. coli* cells. After grown in

LB at 37 °C, expression was induced with 1mM IPTG at A_{600} of 0.6 and moved to 22 °C shaker for 36 -48 hrs.

Next, we show the stimulated emission imaging of lacZ gene expression in live *E. coli* cells. lacZ has been used as a classic reporter gene in various prokaryotic and eukaryotic cells (Miller, 1972). Its protein product, *β*-galactosidase, catalyzes the glycosidic linkage cleavage of X-gal, a popular chromogenic substrate to form a bluish product. Traditionally, the X-gal hydrolysis product has to accumulate in sufficient amount for its blue color to be visible (Miller, 1972).

Wild-type *E. coli* cells are incubated with 50 μM X-gal solution in 37ºC for 30 min, and then concentrated and sandwiched between two No. 1 coverslips. No inducer for lacZ gene is added.

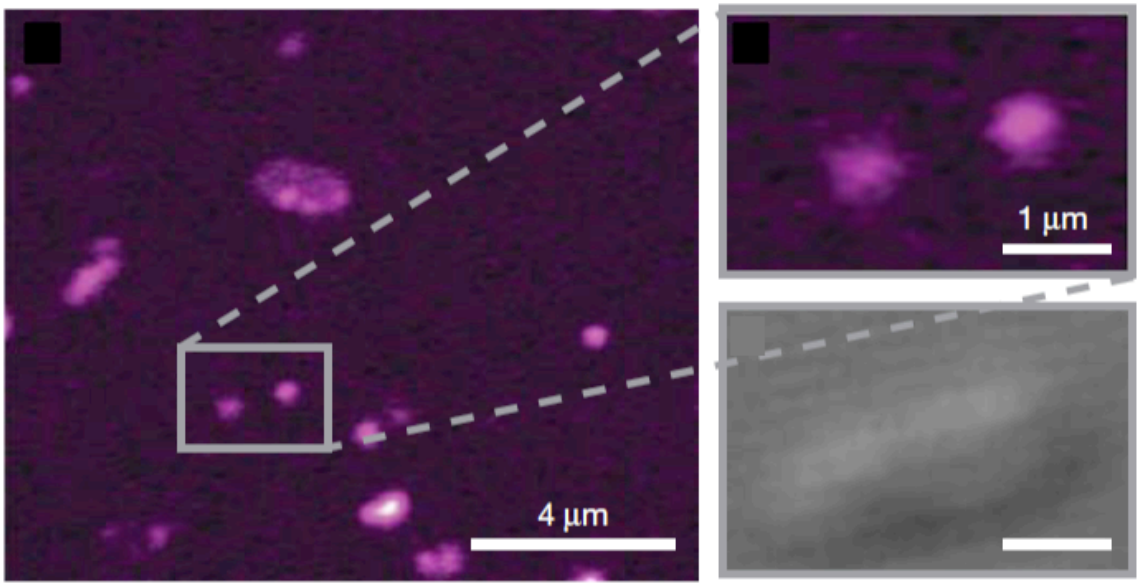


Figure 2.11 Imaging non-fluorescent chromoproteins by stimulated emission. Imaging distributions of cytoplasmic chromoproteins gtCP (a) and cjBlue (c) in live E. coli cells, respectively, by stimulated emission microscopy, overlapped with the corresponding wide-field transmission images. (b) shows the fluorescence yield for gtCP is low, therefore the image quality is low due to stray light and high gain of the PMT.

With the significant improvement of the detection sensitivity of stimulated emission, the basal level lacZ gene expression in the absence of inducer can now be monitored (Figure 2.10). The inhomogeneous distribution of X-gal hydrolysis product inside individual cells is consistent with the fact that this product is insoluble and tends to form localized precipitates. In contrast, the corresponding transmission image shows no signs of any color within the cell. We note that an assay using a fluorogenic substrate has been recently developed (Cai et al., 2006), but it requires a microfluidic container to enclose individual cells because the hydrolysis product

is quickly pumped out by the cell. Hence, stimulated emission microscopy allows monitoring lacZ reporter gene activity with ease and with superb sensitivity.

2.7 Imaging the Distribution of a Drug for Photodynamic Therapy

In this section, we show another application of the stimulated emission microscopy. Here we monitor the transdermal delivery of non-fluorescent drug with intrinsic 3D optical sectioning. Specifically, we show the mapping of a cationic thiazine dye toluidine blue O (TBO) at both the cellular and tissue levels. Having a selective affinity for cancer cells *in vivo*, TBO is an actively explored photosensitizer in photodynamic therapy (Chelvanayagam and Beazley, 1997; Tremblay, 2002). Subcellular localization of photosensitizers is crucial since it influences both the level and the kinetics of inducing apoptosis. However, it is difficult to image the true distribution of TBO, because its fluorescence is quenched when bound to tissue substrates and only the non-specific stain residue in the tissue retains the native fluorescence (Chelvanayagam and Beazley, 1997). Being free of the complication and artifact from fluorescence contrast, stimulated emission microscopy is an ideal method of imaging the drug distribution with high fidelity.

Toluidine blue O (TBO) is used as purchased (Sigma Aldrich). Human embryonic kidney (HEK) 293 cell line was obtained from American Type Culture Collection (ATCC, Rockville),

HEK 293 cells are maintained in DMEM (ATCC) supplemented with 10% fetal bovine serum (ATCC) at 37 ºC in a humidified 5% CO_2 air incubator. Cells are cultured on uncoated glass bottom dishes (P35G-1.0-14-C, MatTek Cooperation). The image is taken one hour after incubating the cells with 10 μM TBO/PBS solution. The stimulated emission image of TBO inside the cancer cells after incubation clearly shows its local accumulation inside the cytoplasm (Figure 2.12).

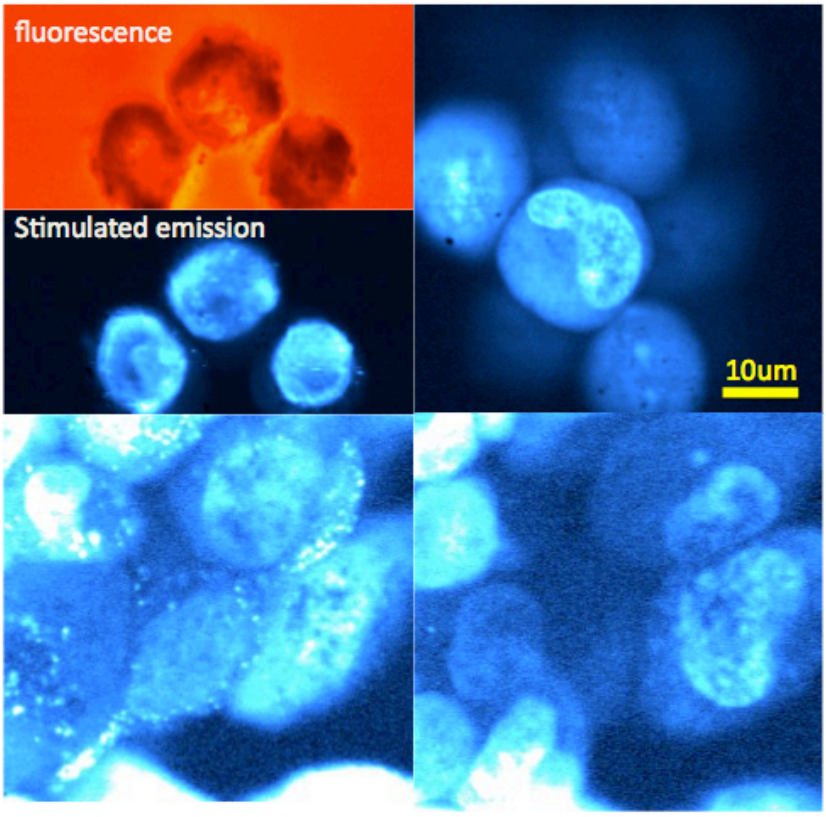


Figure 2.12 Imaging the Subcellular distribution of TBO, a drug widely used for photodynamic therapy by stimulated emission. The blue images present the distribution imaged by stimulated emission, compared with the fluorescence image showing quench of the fluorescence of TBO when binding to the substrate.

Here we show the distribution of TBO in tissue. Mouse skin tissue from wild-type white mice is obtained from Harvard Mouse Facility. Thin ear is harvested for drug incubation immediately after sacrificing the mouse. Approximately 25 μl of a 10 μM TBO/PBS solution is pipetted onto a 5X5 mm piece of skin surface, and the tissue is then incubated at 37 ºC and saturating humidity for one hour. The whole ear tissue is then placed between two No. 1 coverslips for imaging. When topically applied to skin tissue, being hydrophilic and water soluble, TBO is enriched in the center of the protein phase of the polygonal *stratum corneum* cells rather than in the intercellular space which is in lipid phase (Figure 2.13). At a 20 μm deeper depth, TBO displays a rich subcellular distribution in the cytoplasm of viable epidermis where cellular proliferation actively takes place. These imaging results support the hydrophilic delivery pathway as well as the recent hypothesis of TBO binding to cytoplasmic RNA to initiate apoptosis (Tremblay, 2002).

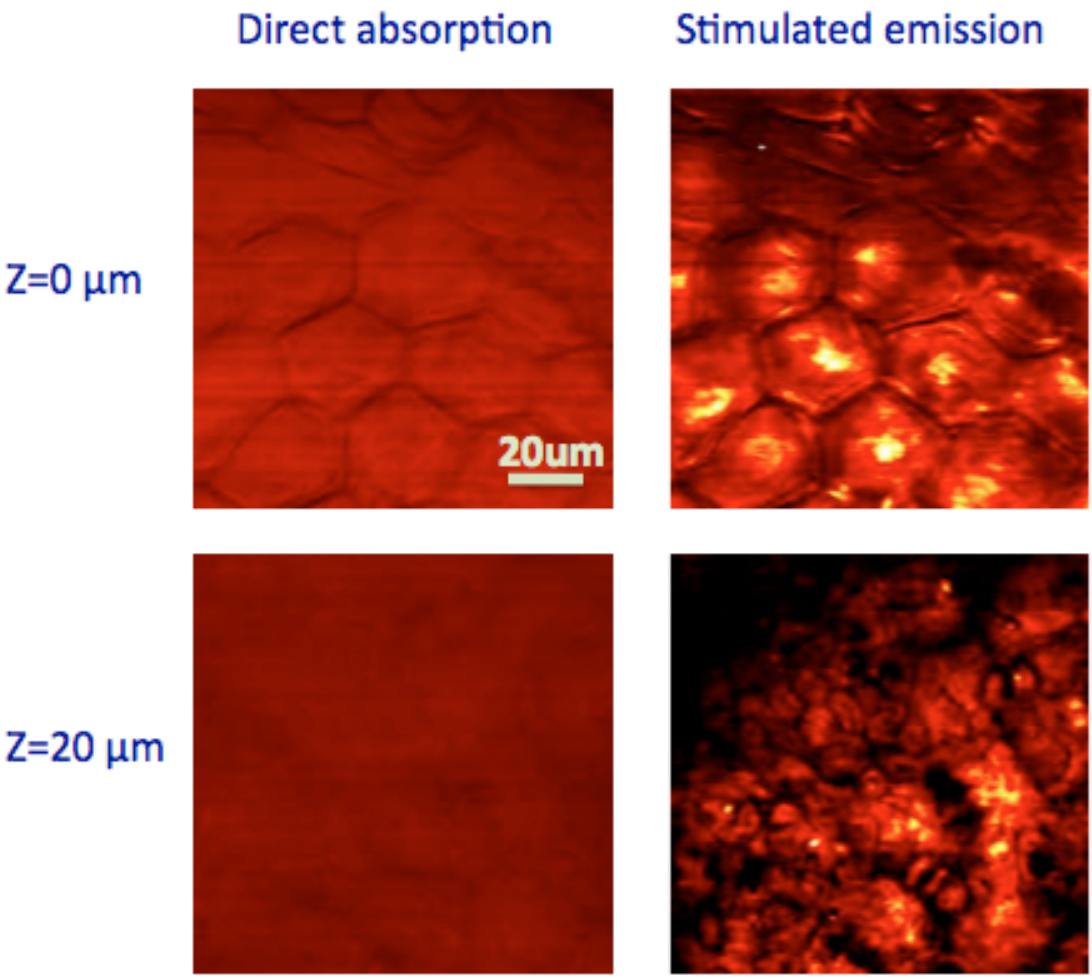


Figure 2.13 Transdermal drug distribution in three-dimension and microvascular imaging. Here shows the drug delivery of TBO to the same area of freshly cut mouse ear skin at two different depths, 30 minutes after topical application of 10uM TBO/PBS solution. At the surface layer of epidermis, TBO accumulates in the protein phase of the polygonal cells rather than in the lipid-rich intercellular space. At the layer of epidermis, a rich TBO distribution following the Subcellular cytoplasm of nucleated basal keratinocytes is shown.

These images support the 'hydrophilic path' as a main pathway for transdermal drug delivery of TBO. Stimulated emission microscopy offers a new approach for studying pharmacokinetics *in situ*.

2.8 Label-Free Microvascular Imaging

Finally, we demonstrate label-free microvascular imaging based on endogenous contrast from non-fluorescent hemoglobin. Blood vessel structure and hemodynamics play a major role in many biomedical processes such as angiogenesis in tumors (McDonald and Choyke, 2003) and cerebral oxygen delivery in brain (Grinvald et al., 1986; Kleinfeld et al., 1998). However, established techniques such as MRI, CT, PET, ultrasound, confocal and two-photon fluorescence microscopy either lack the resolution needed for resolve individual micro-capillaries, or require invasive procedures or exogenous contrasting agents. Here we perform *ex vivo* stimulated emission imaging of the well-developed vascular network from a nude mouse ear, by exciting the Soret band of hemoglobin through two-photon absorption (Clay et al., 2007) and subsequently stimulating the emission from its *Q* band. As shown in Figure 2.14, the single capillaries (~5 μm in diameter) could be clearly identified by the individual blood cells lining within. The spatial interplay between the blood vessels and sebaceous glands is also apparent as the capillaries branch and loop around the tissue structures. Further

imaging of blood oxygenation level by distinguishing oxy- and deoxy- hemoglobin would help to address broad physiological and pathological problems (Wang, 2000).

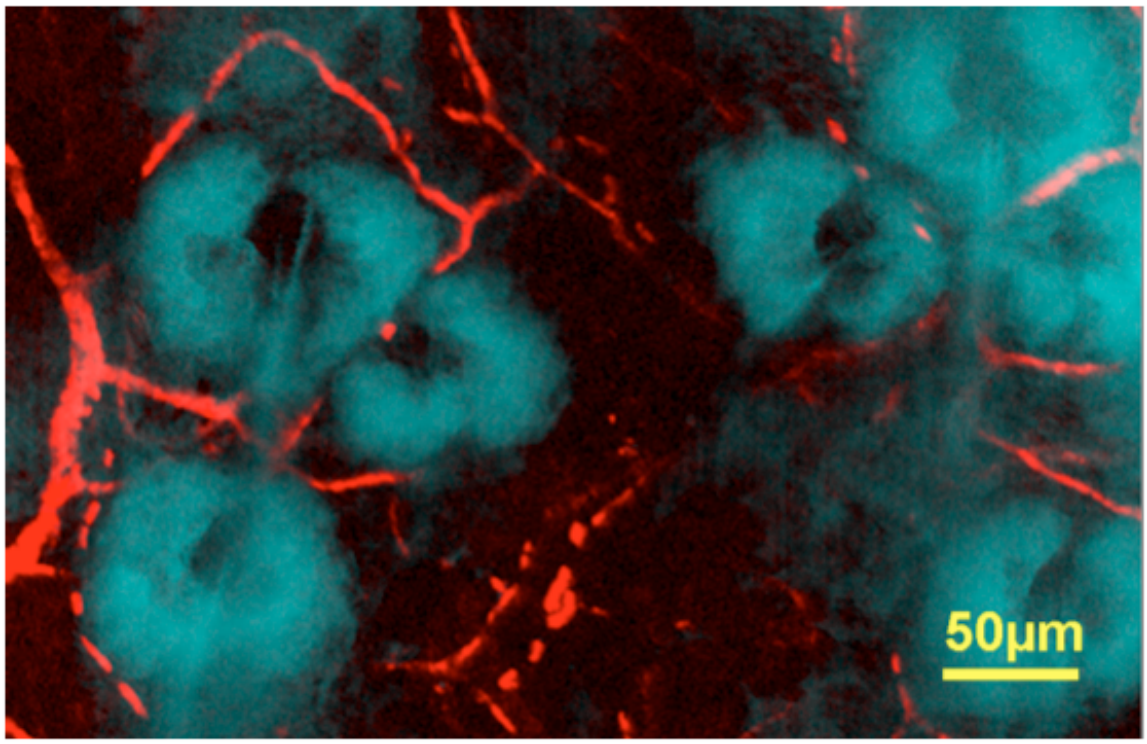


Figure 2.14 Ex vivo imaging of microvasculature network of a mouse ear based on endogenous hemoglobin contrast. The stimulated emission image (red channel, maximum intensity project) shows the blood vessel network surrounding sebaceous glands (cyan channel, simultaneously recorded by confocal reflectance). Individual red blood cells can be resolved within a single capillary. 830nm and 600nm are used for two photon excitation of the Soret band and one-photon stimulated emission of the Q band of hemoglobin, respectively. Pulse widths of both excitation and stimulation beams are about 0.2ps with a ~0.2ps time delay between them.

To summarize, stimulated emission microscopy allows detection and imaging of non-fluorescent chromophores in living cells and tissues with intrinsic 3D optical sectioning and high sensitivity, and extends the repertoire of reporters for biological imaging beyond fluorophores.

References:

Cai, L., Friedman, N., and Xie, X.S. (2006). Stochastic protein expression in individual cells at the single molecule level. Nature *440*, 358–362.

Cantor, C.R., and Schimmel, P.R. (1980). Biophysical Chemistry.

Chan, M.C.Y. (2006). Structural characterization of a blue chromoprotein and its yellow mutant from the sea anemone Cnidopus japonicus. J. Biol. Chem. *281*, 37813–37819.

Chelvanayagam, D.K., and Beazley, L.D. (1997). Toluidine blue-O is a Nissl bright-field counterstain for lipophilic fluorescent tracers Di-ASP, DiI and DiO. J. Neurosci. Methods *72*, 49–55.

Clay, G.O., Schaffer, C.B., and Kleinfeld, D. (2007). Large two-photon absorptivity of hemoglobin in the infrared range of 780-880 nm. J. Chem. Phys. *126*, 025102.

Denk, W., Strickler, J.H., and Webb, W.W. (1990). Two-photon laser scanning fluorescence microscopy. Science *248*, 73–76.

Dong, C.Y., So, P.T., French, T., and Gratton, E. (1995). Fluorescence lifetime imaging by asynchronous pump-probe microscopy. Biophys. J. *69*, 2234–2242.

Du, H. (1998). PhotochemCAD: A computer-aided design and research tool in photochemistry. Photochem. Photobiol. *68*, 141–142.

Einstein, A. (1917). On the quantum theory of radiation. Phys. Z *18*, 121–128.

Evans, C.L., and Xie, X.S. (2008). Coherent anti-Stokes Raman scattering microscopy: chemical imaging for biology and medicine. Annu. Rev. Anal. Chem. *1*, 883–909.

Freudiger, C.W., Min, W., Saar, B.G., Lu, S., Holtom, G.R., He, C., Tsai, J.C., Kang, J.X., and Xie, X.S. (2008). Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy. Science *322*, 1857–1861.

Fu, D. (2007). High-resolution in vivo imaging of blood vessels without labeling. Opt. Lett. *32*, 2641–2643.

Grinvald, A., Lieke, E., Frostig, R.D., Gilbert, C.D., and Wiesel, T.N. (1986). Functional architecture of cortex revealed by optical imaging of intrinsic signals. Nature *324*, 361–364.

Gurskaya, N.G. (2001). GFP-like chromoproteins as a source of far-red fluorescent proteins. FEBS Lett. *507*, 16–20.

Hamilton, C.E., Kinsey, J.L., and Field, R.W. (1986). Stimulated emission pumping: new methods in spectroscopy and molecular dynamics. Annu. Rev. Phys. Chem. *37*, 493–524.

Hell, S.W., and Wichmann, J. (1994). Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Optics Letters *19*, 780–782.

Kleinfeld, D., Mitra, P.P., Helmchen, F., and Denk, W. (1998). Fluctuations and stimulus-induced changes in blood flow observed in individual capillaries in layers 2 through 4 of rat neocortex. Proc. Natl Acad. Sci. USA *95*, 15741–15746.

Lakowicz, J.R. (2006). Principles of Fluorescence Spectroscopy (Springer).

McDonald, D.M., and Choyke, P.L. (2003). Imaging of angiogenesis: from microscope to clinic. Nature Med. *9*, 713–725.

Miller, J.H. (1972). Experiments in Molecular Genetics.

Min, W., Lu, S., Chong, S., Roy, R., Holtom, G.R., and Xie, X.S. (2009). Imaging chromophores with undetectable fluorescence by stimulated emission microscopy. Nature *461*, 1105–1109.

Moerner, W.E., and Kador, L. (1989). Optical detection and spectroscopy of single molecules in a solid. Phys. Rev. Lett. *62*, 2535–2538.

Seigman, A.E. (1986). Laser.

Tremblay, J.F. (2002). Photodynamic therapy with toluidine blue in Jurkat cells: cytotoxicity, subcellular localization and apoptosis induction. Photochem. Photobiol. Sci. *1*, 852–856.

Turro, N.J. (1991). Modern Molecular Photochemistry (University Science Books).

Wang, W. (2000). Femtosecond multicolor pump-probe spectroscopy of ferrous cytochrome c. J. Phys. Chem. B *104*, 10789–10801.

Ye, J., Ma, L.S., and Hall, J.L. (1998). Ultrasensitive detections in atomic and molecular physics: demonstration in molecular overtone spectroscopy. J. Opt. Soc. Am. B *15*, 6–15.

Zhang, J., Campbell, R.E., Ting, A.Y., and Tsien, R.Y. (2002). Creating new fluorescent probes for cell biology. Nature Rev. Mol. Biol. *3*, 906–918.

Chapter 3

Two-Photon Excited Photothermal Lens Microscopy

3.1 Introduction

In Chapter Two, we introduced using absorption and stimulated emission as a contrast mechanism for imaging. Although with high sensitivity, the complex experimental setup compromises the general usage of such microscopic technique. In this chapter, we continue the exploration of using absorption as a contrast mechanism for high-sensitivity 3-D imaging.

As described before, absorption results in intensity loss, which is difficult to be detected in the microscopic setting due to high background and lack of 3-D information. Here we try to detect the ‘secondary’ effect of light absorption, that is, the generation of heat. Light as

particles contain energy, and the energy transfer from light to matter ultimately results in heat dissipation, which can be detected using another beam of light with a different color.

To detect the photothermal effect with 3-D resolution, we employ the two-photon excited photothermal effect as a contrast mechanism to map heme proteins distribution. Particularly, both a thermal lens scheme and a high-frequency modulation are utilized to enhance the signal-to-noise ratio of the detection. We demonstrate label-free imaging of individual red blood cells, mitochondria in live mammalian cells, and the micro-vascular networks in mouse ear tissue and in a zebrafish gill.

Heme proteins, such as hemoglobins and cytochromes, are important biological molecules in most living organisms, as they participate in crucial processes such as electron and oxygen transport and apoptosis (Dawson, 1988)(Choi et al., 1996). To be able to image the distribution of these proteins with high sensitivity and selectivity could greatly facilitate biomedical studies such as tumor angiogenesis and apoptosis signaling (McDonald and Choyke, 2003; Jiang and Wang, 2004). In particular, medical applications such as blood vessel imaging require three-dimensional and high spatial resolution, ideally down to single capillaries. Image contrast generated by endogenous means would be preferable to exogenous contrast agents (Zhang et al., 2006). However, although heme proteins strongly absorb visible

light in their Soret and *Q* bands, they exhibit extremely weak fluorescence quantum yields ($<10^{-5}$) (Champion and Perreault, 1981; Jimenez and Romesberg, 2002). Therefore, developing a label-free non-fluorescence optical imaging technique to visualize these chromophores in their natural physiological environment is both a rewarding and a challenging endeavor. The descriptions in this chapter are based on a previously published work (Lu et al., 2010).

3.2 Photothermal Lensing Effect

The photothermal lensing effect was previously reported and its microscopic application relies on the detection of local heating generated by optical absorption of molecules (Bialkowski, 1995; Tokeshi et al., 2001; Boyer et al., 2002; Cognet et al., 2003; Brusnichkin et al., 2007). Photothermal microscopy is an emerging technique to detect absorbing microscopic objects, and is particularly suitable for imaging metal particles (Boyer et al., 2002; Cognet et al., 2003; Brusnichkin et al., 2007) and it may overcome the difficulty of the poor sensitivity of direct absorption measurements.

To retain the 3-D resolution capability for photothermal microscopy, we explore the usage of two photon excited photothermal effect as a contrast mechanism to image heme proteins with high sensitivity. Ultrafast spectroscopy experiments have shown that heme proteins have

extremely fast internal conversion rates upon photo-excitation and therefore exhibits ultra-short excited state lifetime (<50 femtoseconds) of Soret band (Champion and Perreault, 1981; Jimenez and Romesberg, 2002). The accompanied high conversion efficiency (close to 100%) from optical absorption to heat dissipation suggests that the photothermal effect could be an ideal contrast mechanism for imaging heme proteins.

3.3 Instrumentation

In all reported photothermal imaging schemes, one-photon linear absorption of the laser beam is used to induce local heating, which is then read out by a second probe beam. In contrast, we use two-photon nonlinear excitation by a near infrared (NIR) ultrafast laser as the heating source for heme proteins. This excitation scheme offers a number of different useful features. First, two photon absorption confines the absorption to the exact focal region instead of along the entire beam path, as a consequence of the nonlinear intensity dependence. This is analogous to using two photon excited fluorescence for three dimensional imaging (Denk et al., 1990). Therefore, as shown in Figure 3.1, compared to the conventional case using one-photon excitation, the resulting two-photon excited thermal gradient is less “dilute” and much more concentrated along the z- axis, which would enhance the readout signal generated by the probe beam.

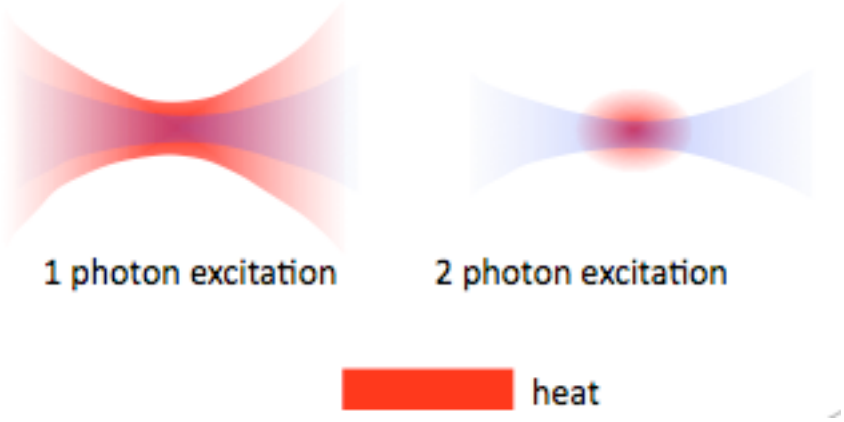


Figure 3.1: Comparison of heat generation by one-photon and two-photon absorption processes. The two-photon absorption is more concentrated in the light focus, therefore it enables photothermal imaging with 3-D sectioning.

The second important feature for two-photon excitation is that, compared to one-photon visible excitation, NIR light has much deeper penetration depth in scattering tissue and causes much less damage on biological samples. Specifically for heme protein, the intense Soret band (~415nm) of heme proteins has recently been shown to exhibit a large and specific two photon absorption cross section near 830nm (Clay et al., 2007).These desirable features of two photon excited photothermal detection enable imaging heme proteins in living cells and in highly scattering tissues with superb sensitivity.

The schematic of the laser scanning microscope is depicted in Figure 3.2. A near IR laser beam at 830nm with repetition rate of 76MHz pulse train (~200 femtosecond pulse width)

serves as the excitation beam, while a stable continuous wave light at 785nm serves as the probe beam. After collinearly combination by a dichroic mirror, these two beams are focused coaxially into the specimen by a microscope objective. The forward propagating beam is collected by a condenser. The probe beam is collected by the condenser-lens pair, spectrally cleaned by a filter, and focused onto a large- area Si photodiode.

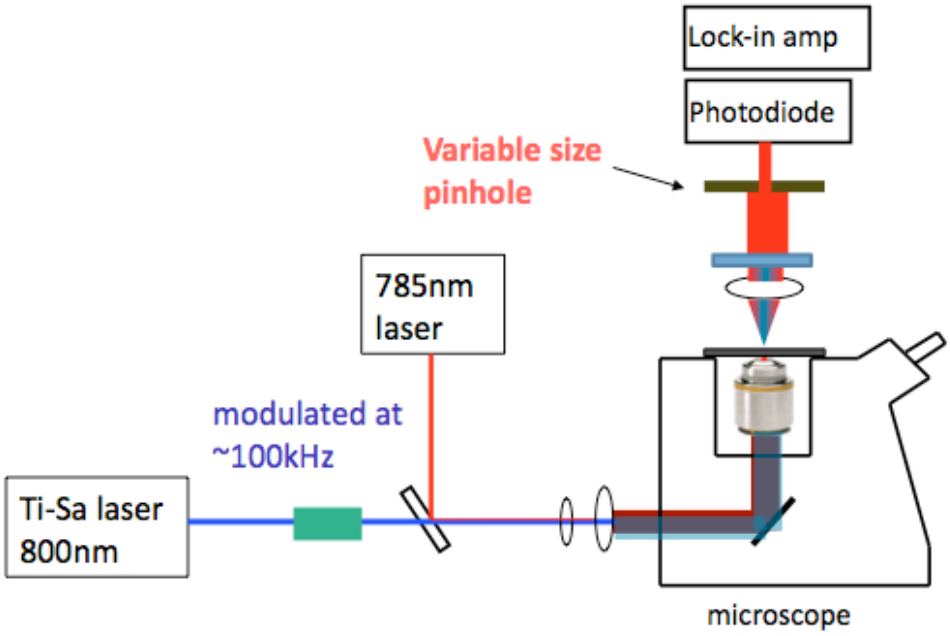


Figure 3.2: Instrumentation setup of a two-photon photothermal lens microscope. The incident ultrafast excitation (~200fs at 830nm) and continuous wave probe beams at 785nm are spatially overlapped and focused onto the sample. A modulator switches the intensity of the excitation beam on-and-off at ~100kHz. While the collinear excitation and probe beams are raster scanned across the sample, the spectrally filtered probe beam is collected by the condenser-lens pair, de-scanned onto an iris diaphragm with adjustable position and aperture, and is refocused onto a large-area photodiode, and demodulated by a lock-in amplifier to create the image contrast.

The detailed instrument setup for the two photon photothermal lens microscope is as follows. To generate 200fs 830nm laser beams, a high power Yb laser was modified to operate at 75MHz. The homemade Yb laser can produce pulses ranging from 100fs to 1ps, at power levels of 6 to 11W at a wavelength of 1040nm. More than 60% conversion to the second harmonic generation at 520nm is routinely achieved with an angel tuned LBO crystal at room temperature (theta=90 degrees, phi=13 degrees, 4mm long, Casix). The doubled 520nm is then used to pump a homemade optical parametric oscillator (OPO), which uses a temperature tuned 6mm long LBO crystal. By changing the temperature of the crystal, the signal wave is tunable from 680nm to 1000nm, and the idler wave is available at wavelengths from 1080nm to more than 2000nm. In the current work, 300mW of 200fs 830nm beam is generated from the OPO signal wave, and the power at the focus is reduced to less than 10mW to avoid photodamage. For the current work, the homemade laser and OPO can also be replaced with a commercial ultrafast Ti: Sapphire laser. The femtosecond excitation beam and the CW probe laser beam (785nm, Sacher Lasertechnik, TEC510-785) are spatially overlapped with a dichroic mirror. The excitation beam is modulated by an acoustic-optical modulator (AOM) (model 3080-122, Crystal technology) at 100kHz which is driven by a square-wave function generator.

Excitation and probe beams are coupled into a modified laser scanning inverted microscope (IX71, FV300, Olympus). The beam size is matched to fill the back-aperture of objective. A

60X 1.2 N.A long-working distance objective (UPlanSApo, water, Olympus) is used for excitation for all cell images (red blood cells and cytochrome imaging) ; A 20X 0.75 N.A objective (UPlanSApo, air, Olympus) is used for excitation for all tissue. A 20X 0.95 N.A. long-working distance objective (XLUMPlanFI, water, Olympus) is used as a condenser. Another lens is used to image the scanning beams onto a silicon amplified photodiode (PDA36A, Thorlabs) to avoid beam movement during laser scanning. Two high OD low pass filters (3RD800SP, Omega) are used together to block the excitation beam completely and only transmit the probe beam. An iris diaphragm with adjustable aperture is mounted on a two dimensional translational stage and is placed in the probe beam path.

The output of the photodiode is low-pass filtered (DC-1.9MHz, Mini-circuits) to suppress the strong signal at the pulsing repetition rate (76MHz), and is terminated into 50Ω. A high-frequency lock-in amplifier (SR844, Stanford Research) is used to demodulate the stimulated emission signal. The analog on phase component x-output of the lock-in amplifier is fed into the A/D converter of the microscope input. The time constant is set for 100us for the imaging experiments. All images are collected with 200μs pixel dwell time, which takes about 13 seconds for a 256×256 pixels image.

3.4 Characterization of the Two-Photon Photothermal Signal

A thermal lens detection scheme, which is among the most sensitive methods of the family of photothermal spectroscopy approaches (Bialkowski, 1995; Tokeshi et al., 2001) is adopted in our method to enhance the sensitivity of probe beam. We wish to detect the refractive index gradient generated by the excitation beam. To do so, an iris diaphragm with an adjustable aperture size and position is installed in front of the detector. Only the central portion of the probe beam instead of the entire beam is allowed to pass through the iris to be detected. The final two photon excited photothermal lens signal can be estimated as:

$$S \propto \frac{I_{ex}^2 \cdot I_{pr} \cdot \sigma_{2-p} \cdot [c] \cdot \eta_H}{\lambda_{pr} \cdot \kappa} \left(\frac{dn}{dT} \right)_P \qquad (3.1)$$

I_{ex} and I_{pr} are the intensity of the excitation beam and the probe beam, respectively, σ_{2-p} is the two-photon absorption cross section, $[c]$ is the sample concentration, η_H is the yield of heat dissipation, λ_{pr} is the wavelength of the probe beam, κ is the thermal conductivity, and $(dn/dT)_P$ is the refractive index temperature coefficient at constant pressure.

To significantly reduce the detection noise, we implement a high-frequency (> 50 kHz) phase-sensitive detection scheme. By doing this, the probe laser intensity fluctuation, which primarily occurs at low frequency (DC to 10 kHz), can be circumvented, as has been previously applied in other spectroscopic and microscopic techniques (Freudiger et al., 2008; Min et al., 2009). In the scheme shown in Figure 3.3, the intensity of the excitation beam is

modulated by an acoustic-optical modulator, creating a modulation of the photothermal signal at the same frequency. Such an induced modulation signal can then be sensitively extracted by a lock-in amplifier referenced to this modulation frequency. In this way, the photothermal signal is detected against a vanishing laser noise, offering superb sensitivity.

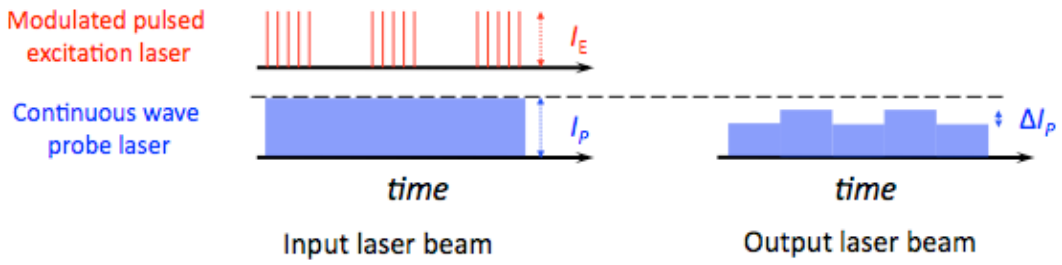


Figure 3.3: Modulation of the excitation laser causes the in-phase change of the intensity of the probe laser. Such intensity modulation can be picked up by a lock-in amplifier.

To verify the image contrast is indeed the result of the two-photon photothermal effect, we probe the photothermal signal of 100uM hemoglobin solution under different pump and probe laser power (Figure 3.4).

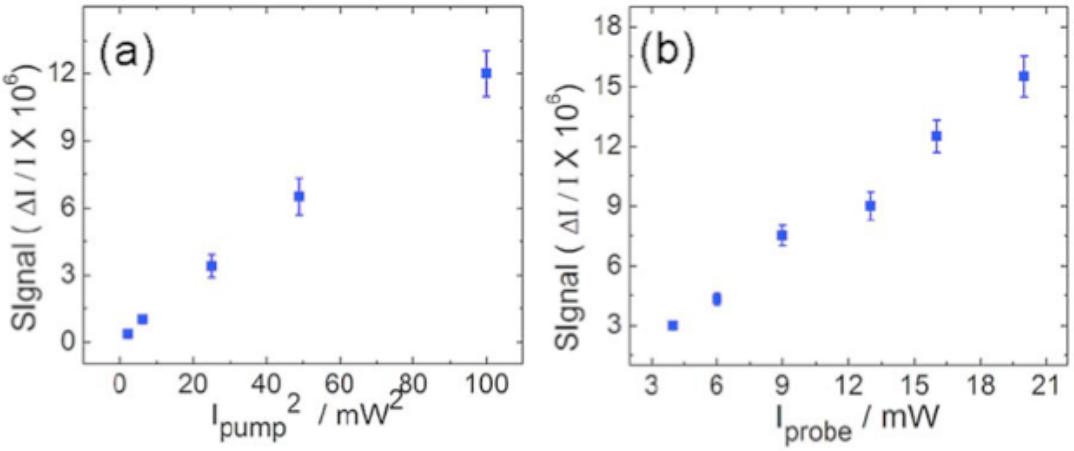


Figure 3.4: Dependence of photothermal signal on the pump and the probe laser power in 100uM hemoglobin solution

As is expected, the overall signal scales quadratically with the power of the pump laser and linearly with the power of the probe laser, which indicates the detected signal is indeed due to two photon photothermal effect.

The accumulation of heat due to photoabsorption changes the refractive index of the material, and thereby modulate the property of the propagation such as the divergence of the probe beam. However, the total intensity of the probe beam should not change due to modulation of refractive index. Therefore, we expect no signal if we collect all the laser power of the probe beam. Such effect is shown in Figure 3.5. We placed an iris in front of the detector, when the iris diaphragm is fully open to collect all the intensity of the forward propagating probe beam,

the photothermal image, is essentially blank. However, when the iris is partially closed to block the peripheral portion of the probe beam, the corresponding image clearly reveals individual red blood cells.

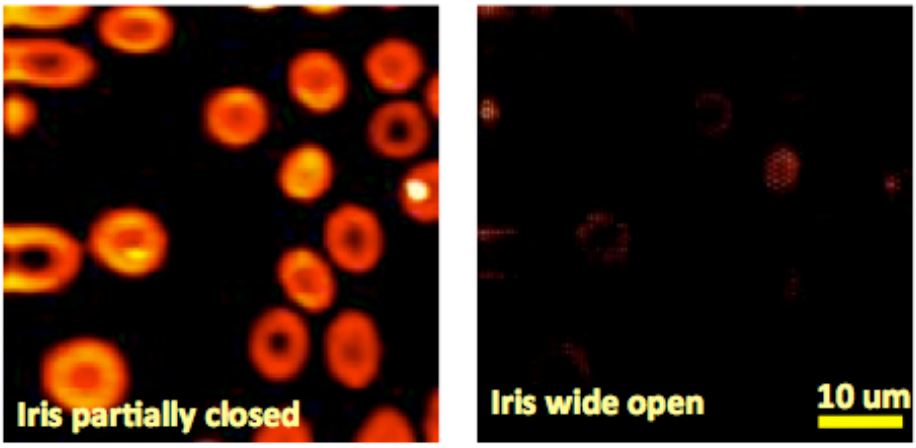


Figure 3.5: photothermal images with iris partially closed and iris widely open. Shown here are individual red blood cells freshly prepared on a microscope slide.

The signal has a sensitive dependence on the collection aperture size (shown in Figure 3.6), which is a manifestation of the underlying thermal lens effect.

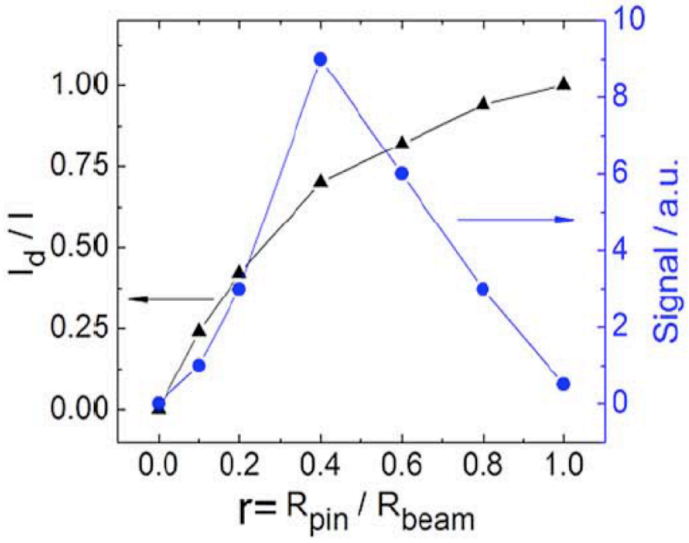


Figure 3.6: photothermal images with iris partially closed and iris widely open. Shown here are individual red blood cells freshly prepared on a microscope slide.

The Modulation frequency of the phase sensitive detection is another crucial parameter in photothermal microscopy. As shown in Figure 3.7, the absolute signal strength increases as the modulation frequency is reduced, because it takes time for the thermal gradient to build up due to finite thermal conductivity. However, the laser noise of the probe beam also starts to increase for the slower modulation frequency. In addition, pixel dwell time has to be significantly longer than the modulation period to be able to demodulate reliably for each pixel. Therefore, based on the balancing between the signal size, the noise level and the

imaging speed, we choose ~100 kHz modulation frequency as the optimum for scanning microscopy.

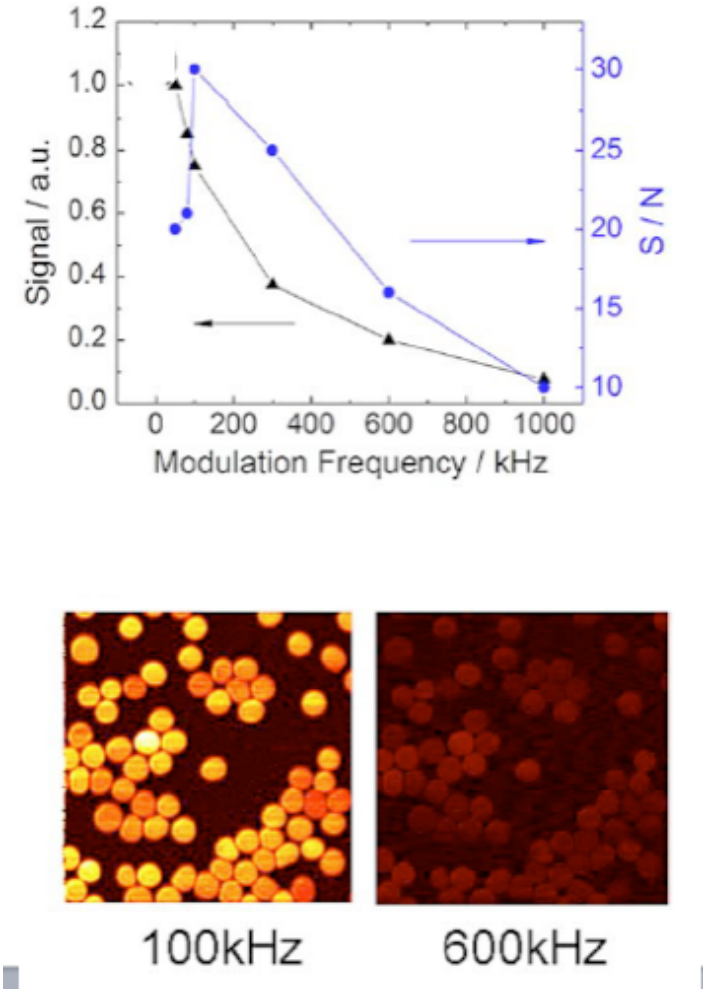


Figure 3.7: Dependence of photothermal signal and signal to noise ratio (S/N) on the modulation frequency from 50kHz to 1MHz. To maintain high signal, low noise and reasonably fast imaging speed, the modulator is set at ~100kHz for all of the following images.

3.5 Bioimaging with Two-Photon Photothermal Lens Microscopy

Here we demonstrate the application of this microscopy in live cell imaging. Human embryonic kidney (HEK) 293 cell line was obtained from American Type Culture Collection (ATCC, Rockville), HEK 293 cells are maintained in DMEM (ATCC) supplemented with 10% fetal bovine serum (ATCC) at 37^{o}C in a humidified 5% CO_2 air incubator. Cells are then plated on uncoated glass bottom dishes (P35G-1.0-14-C, MatTek Cooperation) for imaging.

Mitochondria contain membrane-bound cytochromes which have important roles in electron transport and controlling of apoptosis. Cytochromes are small protein molecules (~12kD) and are difficult to label without affecting their normal physiology. They contain heme group as their key structural components, and therefore can be imaged in this label-free manner. As shown in Figure 3.8, the asymmetric cellular distribution of cytochromes is shown and individual mitochondria can be clearly resolved. This technique could allow study of fusion-fission mitochondrial dynamics in live cells in a label-free manner (Detmer and Chan, 2007).

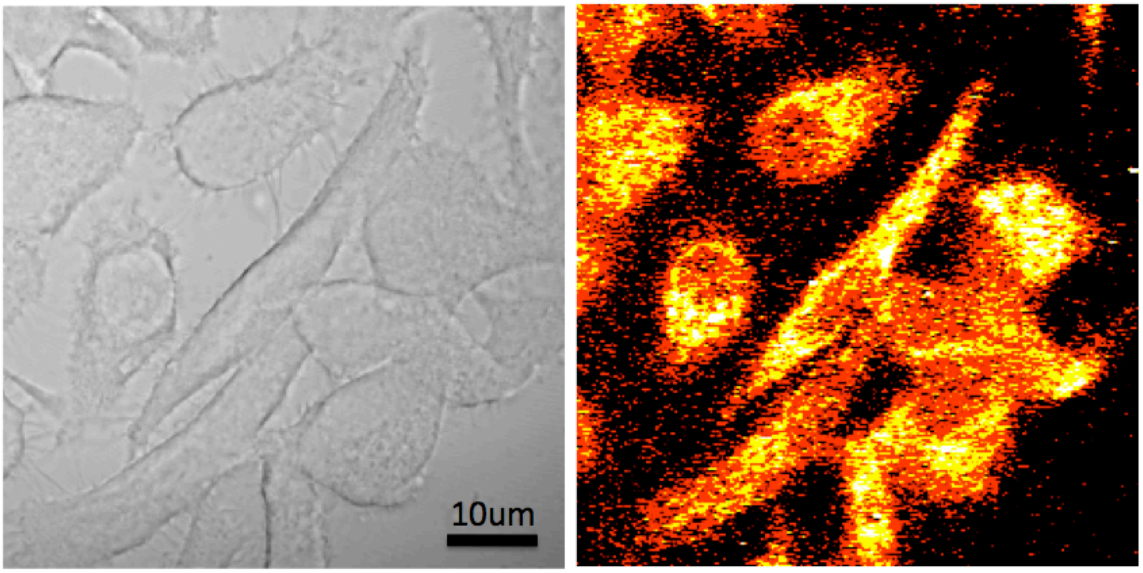


Figure 3.8: Imaging distribution of cytochromes in live HEK-293 cells, in comparison with the transmission image of the probe beam (gray)

Finally, we demonstrate label-free imaging of micro-vascular network based on endogenous contrast from hemoglobin. Structure and hemodynamics of blood vessels play a major role in many biomedical processes such as angiogenesis in tumors (McDonald and Choyke, 2003). However, established techniques such as MRI, CT, PET, ultrasound, and more recently, photoacoustic tomography (Stein et al., 2009), confocal, two-photon microscopy (Schaffer et al., 2006) and fluorescence microendoscopy (Monfared et al., 2006) either lack the spatial resolution needed to resolve individual red blood cells, or require exogenous contrast agents. Figure 3.9 shows the image of vascular network from a mouse ear. Mouse skin tissue from wild-type white mice is obtained from Harvard Mouse Facility. A mouse ear is harvested for

imaging soon after sacrificing the mouse. The mouse ear tissue is then sandwiched between two No.1 coverslips with a drop of water for imaging. The capillaries are seen to wrap around the sebaceous glands shown in the transmission image.

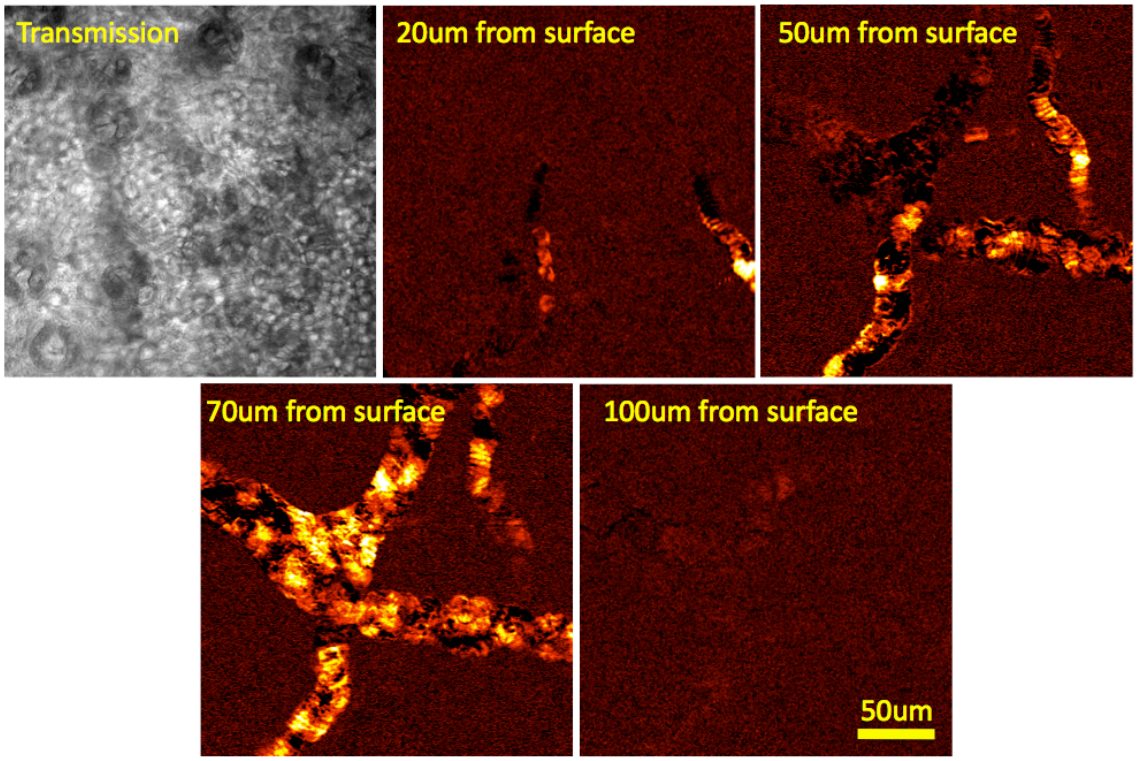


Figure 3.9 Imaging of the micro-vascular network in mouse ear tissue based on hemoglobin contrast. The image shows the blood vessel network surrounding sebaceous glands. Individual blood cells are shown to be lined up within a single capillary (<5μm in diameter). Transmission image is generated by detecting the CW probe beam of the same tissue region.

Furthermore, we can also perform imaging on a whole organism level. Zebrafish, whose larvae are relatively transparent in early development, provides an ideal vertebrate model for

cancer progression and angiogenesis and is readily amenable to genetic and pharmacological screening (Stoletov et al., 2007). Figure 3.10 shows a 3D reconstruction of the blood vessel of a larval zebrafish gill. In contrast with the transmission image, two-photon photothermal contrast allows deep penetration in scattering tissue.

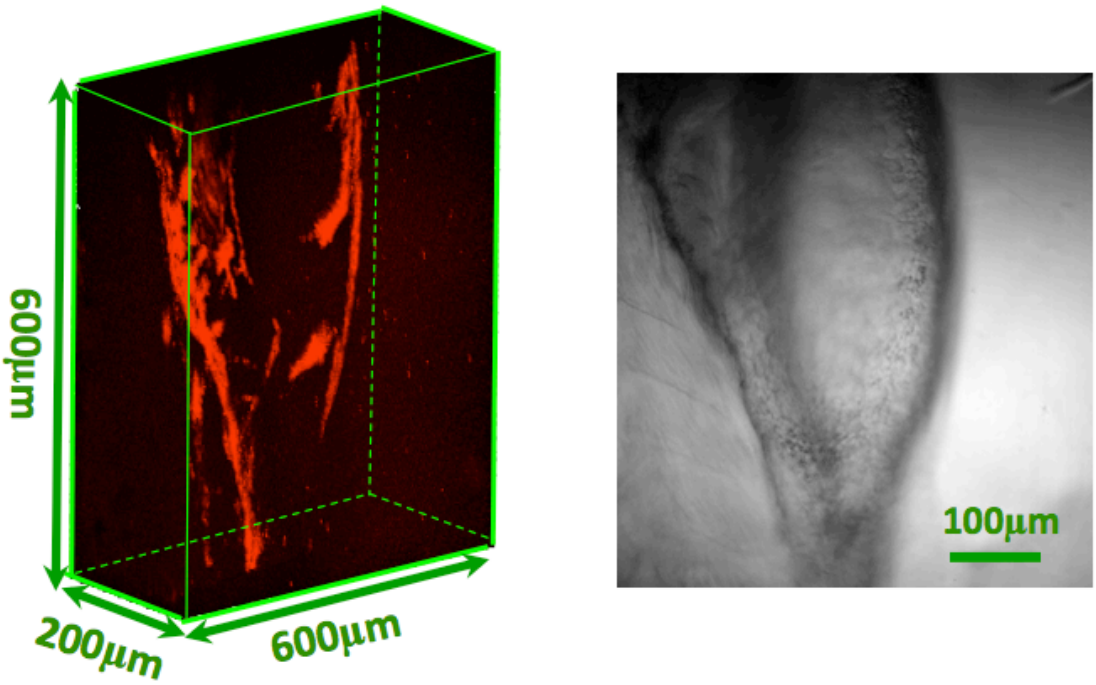


Figure 3.10 3-D reconstruction of blood vessel network in zebrafish gill compared with the transmission image of zebrafish gill by the CW probe beam. In all two-photon excited photothermal imaging experiments, the power of 200fs 830nm excitation beam and 785nm CW probe beam are kept to be ~10mW and ~20mW, respectively, at the focus of the objective, with laser modulation frequency at 100kHz and pixel dwell time ~200µs.

From the perspective of instrumentation, this microscopy can be readily configured from a standard two-photon fluorescence microscope, which is already equipped with the two-photon excitation source. The only additional elements required are a CW probe laser and a modulator/demodulator. We note that other femtosecond pump-probe techniques have recently been employed to image hemoglobin by using two synchronized femtosecond lasers (Fu et al., 2007), which is however technically complex and more expensive.

To summarize, two-photon excited photothermal lens microscopy enables label-free high-resolution imaging of heme protein in live cells, tissues and organisms with intrinsic 3D optical sectioning and high sensitivity. The technique opens up further possibilities of functional imaging of heme proteins such as the oxygenation level of hemoglobin and redox state dynamics of cytochromes, both in cellular and in tissue/organism levels.

References:

Bialkowski, S.E. (1995). Photothermal Spectroscopy Methods for Chemical Analysis (Wiley-Interscience).

Boyer, D., Tamarat, P., Maali, A., Lounis, B., and Orrit, M. (2002). Photothermal Imaging of Nanometer-Sized Metal Particles Among Scatterers. Science *297*, 1160–1163.

Brusnichkin, A.V., Nedosekin, D.A., Proskurnin, M.A., and Zharov, V.P. (2007). Photothermal lens detection of gold nanoparticles: theory and experiments. Appl Spectrosc *61*, 1191–1201.

Champion, P.M., and Perreault, G.J. (1981). Observation and quantitation of light emission from cytochrome c using Soret band laser excitation. The Journal of Chemical Physics *75*, 490.

Choi, A.M., Alam, J., and others (1996). Heme oxygenase-1: function, regulation, and implication of a novel stress-inducible protein in oxidant-induced lung injury. American Journal of Respiratory Cell and Molecular Biology *15*, 9.

Clay, G.O., Schaffer, C.B., and Kleinfeld, D. (2007). Large two-photon absorptivity of hemoglobin in the infrared range of 780-880 nm. J. Chem. Phys. *126*, 025102.

Cognet, L., Tardin, C., Boyer, D., Choquet, D., Tamarat, P., and Lounis, B. (2003). Single metallic nanoparticle imaging for protein detection in cells. PNAS *100*, 11350–11355.

Dawson, J.H. (1988). Probing structure-function relations in heme-containing oxygenases and peroxidases. Science *240*, 433–439.

Denk, W., Strickler, J.H., and Webb, W.W. (1990). Two-photon laser scanning fluorescence microscopy. Science *248*, 73–76.

Detmer, S.A., and Chan, D.C. (2007). Functions and dysfunctions of mitochondrial dynamics. Nature Reviews Molecular Cell Biology *8*, 870–879.

Freudiger, C.W., Min, W., Saar, B.G., Lu, S., Holtom, G.R., He, C., Tsai, J.C., Kang, J.X., and Xie, X.S. (2008). Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy. Science *322*, 1857–1861.

Fu, D., Ye, T., Matthews, T.E., Chen, B.J., Yurtserver, G., and Warren, W.S. (2007). High-resolution in vivo imaging of blood vessels without labeling. Opt. Lett. *32*, 2641–2643.

Jiang, X., and Wang, X. (2004). Cytochrome C-Mediated Apoptosis. Annual Review of Biochemistry *73*, 87–106.

Jimenez, R., and Romesberg, F.E. (2002). Excited State Dynamics and Heterogeneity of Folded and Unfolded States of Cytochrome c. J. Phys. Chem. B *106*, 9172–9180.

Lu, S., Min, W., Chong, S., Holtom, G.R., and Xie, X.S. (2010). Label-free imaging of heme proteins with two-photon excited photothermal lens microscopy. Applied Physics Letters *96*, 113701–113701–3.

McDonald, D.M., and Choyke, P.L. (2003). Imaging of angiogenesis: from microscope to clinic. Nature Med. *9*, 713–725.

Min, W., Lu, S., Chong, S., Roy, R., Holtom, G.R., and Xie, X.S. (2009). Imaging chromophores with undetectable fluorescence by stimulated emission microscopy. Nature *461*, 1105–1109.

Monfared, A., Blevins, N.H., Cheung, E.L.M., Jung, J.C., Popelka, G., and Schnitzer, M.J. (2006). In Vivo Imaging of Mammalian Cochlear Blood Flow Using Fluorescence Microendoscopy. Otology & Neurotology *27*, 144–152.

Schaffer, C.B., Friedman, B., Nishimura, N., Schroeder, L.F., Tsai, P.S., Ebner, F.F., Lyden, P.D., and Kleinfeld, D. (2006). Two-Photon Imaging of Cortical Surface Microvessels Reveals a Robust Redistribution in Blood Flow after Vascular Occlusion. PLoS Biol *4*, e22.

Stein, E.W., Maslov, K., and Wang, L.V. (2009). Noninvasive, in vivo imaging of blood-oxygenation dynamics within the mouse brain using photoacoustic microscopy. Journal of Biomedical Optics *14*, 020502–020502–3.

Stoletov, K., Montel, V., Lester, R.D., Gonias, S.L., and Klemke, R. (2007). High-resolution imaging of the dynamic tumor cell–vascular interface in transparent zebrafish. PNAS *104*, 17406–17411.

Tokeshi, M., Uchida, M., Hibara, A., Sawada, T., and Kitamori, T. (2001). Determination of Subyoctomole Amounts of Nonfluorescent Molecules Using a Thermal Lens Microscope: Subsingle-Molecule Determination. Anal. Chem. *73*, 2112–2116.

Zhang, H.F., Maslov, K., Stoica, G., and Wang, L.V. (2006). Functional photoacoustic microscopy for high-resolution and noninvasive in vivo imaging. Nature Biotechnology *24*, 848–851.

Chapter 4

Near Degenerate Four-Wave Mixing Microscopy

4.1 Summary and Introduction

Fluorescence microscopy has been widely used because of its high sensitivity (Pawley, 2006), but it is limited to fluorescent samples. Hence, various sensitive spectroscopic contrasts have been explored for imaging non-fluorescent species (1996; Kneipp et al., 1997; Boyer et al., 2002; Dudovich et al., 2002; Sfeir et al., 2004; Hiki et al., 2006; Ignatovich and Novotny, 2006; Armani et al., 2007; Fu et al., 2007; Lasne et al., 2007; Li et al., 2008; Evans and Xie, 2008). In the previous chapters, we have introduced two different methods of using modulation transfer to do 3-D imaging based on absorption contrast (Min et al., 2009a; Lu et al., 2010). These methods, however, do require the molecules to have detectable absorption,

preferably in the visible wavelengths of light.

Nonlinear coherent spectroscopies, such as CARS, second harmonic generation (SHG) and third harmonic generation (THG), offer powerful contrast mechanism for sensitive detection and imaging of non-fluorescent molecules (Shen, 1988; Boyd, 2008). First, as in two-photon fluorescence microscopy (Denk et al., 1990), nonlinear coherent spectroscopies offer intrinsic 3D optical sectioning. Second, the nonlinear wave mixing could generate a signal at a color different from that of the incident laser, creating a nearly background-free situation. Third, coherent amplification could occur due to constructive interference among all the nonlinear induced dipoles. However, one or more virtual states are often involved in the underlying optical transition, which limits the maximum nonlinear coherent signal generation. In addition, symmetry and phase matching condition often limit the harmonic generation microscopies. In particular, SHG only works for non-centrosymmetric materials (Hellwarth and Christensen, 1975; Gannaway and Sheppard, 1978). THG only arises from interfaces or small inclusions because of the large phase mismatch associated with Guoy phase shift near focus (Shen, 1988; Squier et al., 1998; Cheng and Xie, 2002, 2004; Barad et al., 2003; Débarre et al., 2006; Boyd, 2008).

In this chapter, we introduce using single-beam near-degenerate four-wave-mixing (ND-FWM) as a contrast mechanism for doing nonlinear optical imagine, by detecting a coherent signal generated at the spectral "edge" of the incident "shaped" broadband femtosecond laser. ND-FWM microscopy allows label-free imaging of live cells and tissues with high sensitivity and spatial resolution. In particular, by achieving a nearly perfect phase matching condition, ND-FWM provides a contrast mechanism different from other nonlinear imaging techniques such as coherent anti-Stokes Raman scattering (CARS) (Dudovich et al., 2002; Evans and Xie, 2008; Li et al., 2008), second (Hellwarth and Christensen, 1975; Gannaway and Sheppard, 1978) and third (Squier et al., 1998; Barad et al., 2003) harmonic generations. We further develop an electronic resonant version of ND-FWM for absorbing but non-fluorescent molecules. Ultrasensitive chromophore detection (~50 molecules) and hemoglobin imaging are demonstrated, by utilizing a fully resonant enhancement of the nonlinear polarization and optical heterodyne detection**.** The descriptions in this chapter are based on a previously published work (Min et al., 2009c)

4.2 Instrumentation

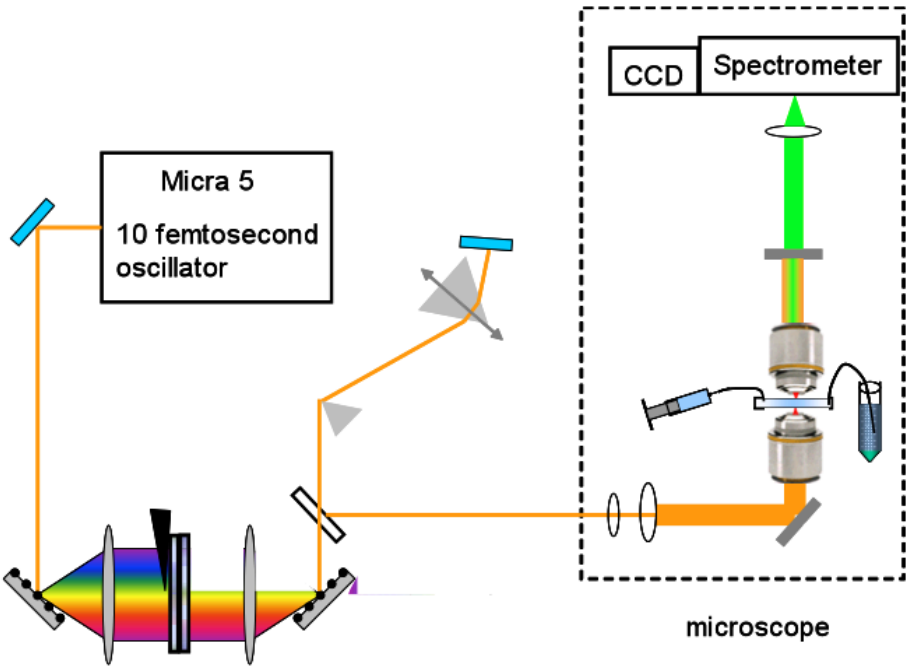


Figure 4.1: Experimental apparatus of ND-FWM spectroscopy and microscopy. The spectra at corresponding stages (after the laser, after the pulse shaper, after the sample, in front of the detector) are illustrated.

As shown in Figure 4.1, the "blue" portion of the broadband Ti-Sapphire laser is sharply blocked by a 4-f pulse-shaper (Weiner, 2000) in its Fourier-domain. After exciting the sample with such a shaped pulse, ND-FWM signal is then generated at new frequencies close to the spectral "cut edge". Therefore, unlike the strictly degenerate FWM, ND-FWM involves similar but not identical input frequencies, and the generated signal can be spectrally separated from the incident light. With a set of high-quality optical filter, the optical detection can be done in a background-free manner.

For detailed experimental setup of a ND-FWM microscope, a 20 fs pulse train centered around 800nm with 80 MHz repetition rate is generated from a commercial Ti:Sapphire oscillator (Micra 5, Coherent Inc.). It is then sent into a 4-f pulse shaper where a sharp razor blade is placed in the spectral plane to physically block the high energy side (<772nm) of the broad laser spectrum. A prism pair pulse compressor is installed to pre-compensate the positive group velocity dispersion introduced by the high N.A. objective. The shaped pulse is then focused by a 1.35 N.A. objective (oil, UPLANSAPO, Olympus) onto the sample (with an averaged power of 0.5~2 mW) in a modified inverted microscope (TE300, Nikon). The sample is raster scanned with a computerized *x*, *y* stage (E-500, Physik Instrumente). The forward going ND-FWM signal after the sample is collected by another objective (0.9 N.A. Zeiss), and filtered by a sharp-edge short-pass filter (770AESP, Omega Optical), and focused onto a red-sensitive PMT (R9110, Hamamatsu) coupled with a low-noise current preamplifier (SR570, Stanford Research).

For measuring the spectrum of ND-FWM signal, The forward collected ND-FWM signal from either glass or neocyanine solution is focused onto the entrance slit of a triple

monochromator (XY, Horiba Jobin Yvon) equipped with a CCD spectroscopy camera (DU920N-BR-DD, Andor). Spectra acquisition time is 100ms.

4.3 Characterization of the ND-FWM Signal

The general FWM arises from a third-order nonlinear interaction of four coherent optical fields in the material. The general form of the induced polarization at frequency ω_4 can be expressed as (Boyd, 2008):

$$P_i^{(3)}(\omega_4 = \pm\omega_1 \pm \omega_2 \pm \omega_3) \propto \sum_{jkl} \sum_{(1,2,3)} \chi_{ijkl}^{(3)}(-\omega_4; \pm\omega_1, \pm\omega_2, \pm\omega_3) E_j(\omega_1) E_k(\omega_2) E_l(\omega_3) \quad (4.1)$$

Equation 4.1 describes a coupling between four waves through the nonlinear susceptibility tensor $\chi_{ijkl}^{(3)}$. The notation $\sum_{(1,2,3)}$ requires the relation $\omega_4 = \pm\omega_1 \pm \omega_2 \pm \omega_3$ to be held in performing the summation. The frequency values of photons that are destroyed in the nonlinear process are written with positive signs, and the created frequencies with negative signs. $\chi_{ijkl}^{(3)}$ for ND-FWM is given by (Boyd, 2008):

$$\chi^{(3)}(-\omega_4; \omega_1, -\omega_2, \omega_3) = \frac{N}{\hbar^3} \mathbf{P}_F \sum_{m,l,n} \frac{\mu_{km}\mu_{ml}\mu_{ln}\mu_{nk}}{(\omega_{mk} - \omega_1 - i\gamma_{mk})(\omega_{lk} - \omega_1 + \omega_2 - i\gamma_{lk})(\omega_{nk} - \omega_4 - i\gamma_{nk})} \quad (4.2)$$

$\mathbf{P}_F$ is the full permutation operator, μ is the transition dipole moment, ω is the energy difference between corresponding energy levels, γ is the homogeneous linewidth of the associated electronic or vibrational transition.

The Spectroscopy of ND-FWM can be well understood as four wave mixing events integrated across a continuously distributed laser spectrum. The measured ND-FWM spectrum exhibits a characteristic decay in frequency from small frequency shift to large shift, with respect to the incident spectrum, as shown in Figure 4.2.

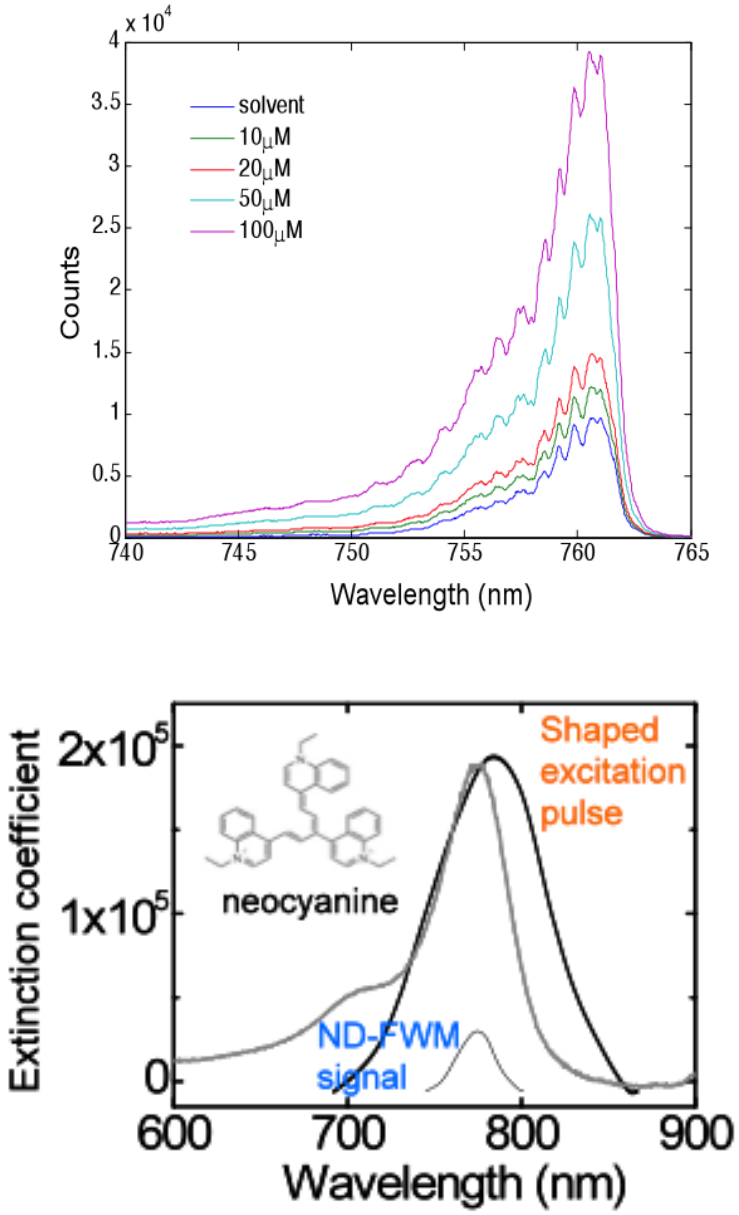


Figure 4.2: The generated ND-FWM spectrum from the solvent (ethanol) and different concentration of neocyanine molecules under tight focus condition (N.A=1.4). The spectrum of the laser output and the absorption spectrum of neocyanine is compared side by side.

Such a decaying spectrum is due to the fact that, given the incident broadband spectrum, there exist more frequency combinations matching $\omega_4 = \omega_1 - \omega_2 + \omega_3$ for smaller frequency shift. Quantitatively, we can describe the spectrum with a continuous integration as

$$I(\omega_4)d\omega_4 \propto \left|P^{(3)}(\omega_4 = \omega_1 - \omega_2 + \omega_3)\right|^2 \propto$$
$$\left|\int_{-\infty}^{\infty} d\omega_1 \int_{-\infty}^{\infty} d\omega_2 \int_{-\infty}^{\infty} d\omega_3 \chi^{(3)}(-\omega_4;\omega_1,-\omega_2,\omega_3)E(\omega_1)E^*(\omega_2)E(\omega_3)\delta(\omega_4 - \omega_1 + \omega_2 - \omega_3)\right|^2 \quad (4.3)$$

in which the delta function ensures the triple integration satisfy $\omega_4 = \omega_1 - \omega_2 + \omega_3$. The experimental measurement agrees well with the theoretical prediction equation 4.3 which is numerically computed. Such a decaying spectrum is precisely the reason why we approach to the small frequency shifts as closely as possible by means of laser spectral shaping.

ND-FWM microscopy provides a bulk-sensitive electronic nonlinear contrast mechanism that is different from CARS, SHG or THG. Coherent signal generation of a general FWM process is only efficient near the phase matching condition (Potma et al., 2000). Moreover, Gouy phase shift (Novotny and Hecht, 2006) has been shown to play an important role in nonlinear coherent microscopy, by generating a phase delay for the excitation field along the axial

direction of the tight focus. A modified wave-vector mismatch has been introduced to account for this:

$$\Delta\mathbf{k}_G = \mathbf{k}_4 - \left[\pm(\mathbf{k}_3 + \delta\mathbf{k}_{3,G}) \pm (\mathbf{k}_2 + \delta\mathbf{k}_{2,G}) \pm (\mathbf{k}_1 + \delta\mathbf{k}_{1,G})\right] \tag{4.4}$$

in which $\delta\mathbf{k}_G$ denotes the contribution from Gouy phase shift of the excitation fields (Cheng and Xie, 2002, 2004). Furthermore, in a forward collinear geometry, Eq. (4.4) can be simplified as:

$$\Delta\mathbf{k}_G = \frac{n_4\omega_4 \mp n_3\omega_3 \mp n_2\omega_2 \mp n_1\omega_1}{c} + \frac{\pi}{2}\left(\pm\frac{1}{\lambda_3} \pm \frac{1}{\lambda_2} \pm \frac{1}{\lambda_1}\right) \tag{4.5}$$

in which n is the refractive index of the material at corresponding frequency ω. ND-FWM offers two distinct mechanisms in optimizing the phase matching condition. First, by having near degeneracy, the refractive index can be regarded as nearly invariant within such a narrow frequency range. As a result, $n_4\omega_4 - n_3\omega_3 + n_2\omega_2 - n_1\omega_1 \approx 0$. Second, $\delta\mathbf{k}_G$ from Guoy phase shift is greatly cancelled by the negative contribution from the conjugated field $E^*(\omega_2)$. This ideal phase matching condition leads to a coherence length longer than the focal depth, and therefore a complete constructive interference within the whole focal volume from a bulk

sample. In contrast, $\delta \mathbf{k}_G$ from three fields in THG add up positively, which results in minimal THG signal from bulk homogeneous medium under the tight-focusing condition (Cheng and Xie, 2002, 2004; Débarre et al., 2006).

ND-FWM microscopy is further characterized in Figure 4.3.

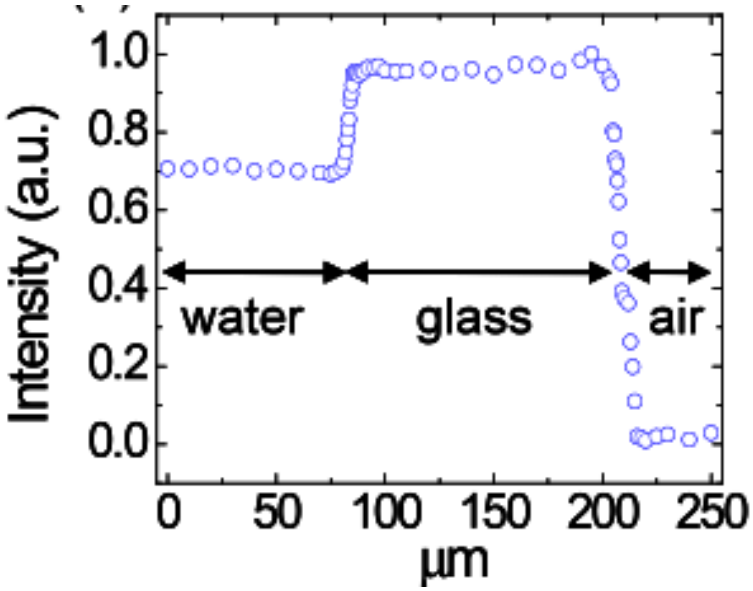


Figure 4.3: The z- scanning ND-FWM intensity profile of the water/glass/air interfaces.

The *z*- scanning profile over water/glass/air proves that this technique is indeed bulk-sensitive instead of interface sensitive. This is predicted by the above theoretical discussion on the nearly perfect phase matching condition of ND-FWM microscopy. In contrast, SHG and THG signals would only arise at the interface but not inside the water or glass. As a simplest

demonstration, Fig. 4.4 shows the image of polystyrene beads (diameter ~ 600nm) dispersed on a glass surface. Meanwhile, a control image of the same area is taken when the excitation pulse width is stretched by tuning the prism pair pulse compressor.

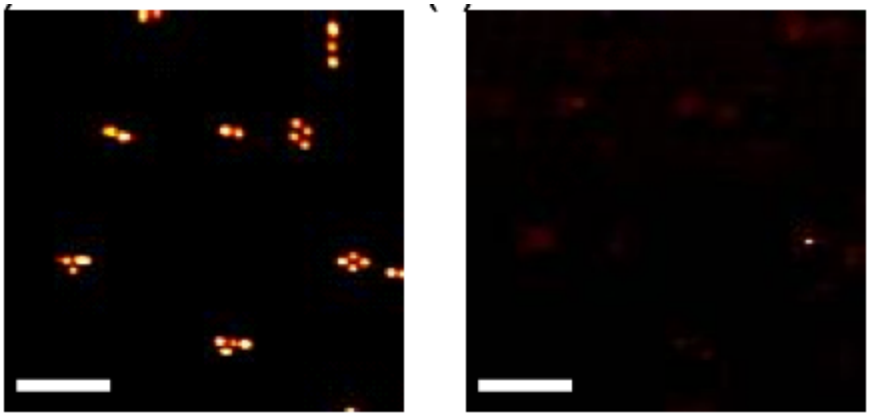


Figure 4.4: Image of polystyrene beads (d~600nm) dispersed on the glass surface, compared with the control image on the same area but excited with stretched excitation pulse width. These images demonstrate the nonlinear nature of the image contrast.

The fact that the image contrast almost vanishes for the much longer pulse width indicates that the signal in Fig 4.4 is indeed from the nonlinear process instead of the linear refractive index variation. The measured lateral FWHM for 50nm bead is about 260nm, which is better than the one-photon confocal resolution of 290nm calculated as $0.61\lambda/1.4$NA (λ~800nm), due to the third order intensity dependence. Visible light (e.g. 500nm) would allow resolution even sharper than 150nm, made possible by the near degeneracy between excitation and detection.

4.4 Bioimaging by ND-FWM

The $\chi^{(3)}_{ijkl}$ for the degenerate FWM process could range from 10^{-8} to 10^{-15} cm^2/W, and it is highly sensitive on the degree of electron delocalization in conjugated systems (Débarre et al., 2006; Boyd, 2008), making ND-FWM a useful contrast mechanism for imaging biological samples without exogenous labels.

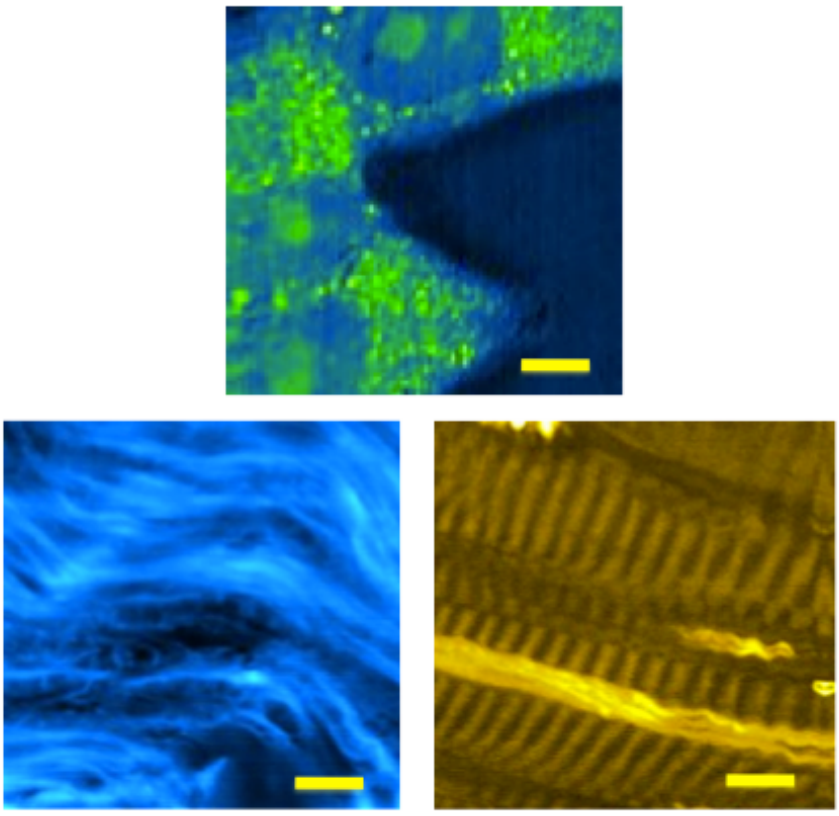


Figure 4.5: Live human lung cancer cells (cultured A549 cell line). Cellular organelles such as nucleus and mitochondria are visible. Tissue morphology is demonstrated by muscle tissue and brain tissue slices, which were freshly cut off from a sacrificed mouse. Scale bar: 10 μm.

The mammalian cell image (Figure 4.5) identifies the sub-cellular organelles such as nucleus and mitochondria. The characteristic fiber morphology as well as cigar-shaped multinucleate cells is clearly visible in the muscle tissue image, and the brain images shows fiber tracts in the corpus callosum.

4.5 Electronic Resonance in ND-FWM

We now explore the electronic resonance version of ND-FWM, which would further enhance its sensitivity and specificity. According to Equation (4.2), a full resonance condition can be achieved when the excitation laser frequency is tuned to resonance with a molecular electronic transition. The resonance enhancement is not only occurring for the excitation frequency but also for the emission/signal frequency, because of the near degeneracy. Furthermore, the low-frequency electronic coupled vibronic states are also resonantly excited by the broadband pulse. Therefore, the triple resonance minimizes all three factors in the denominator of Equation (4.2), thus significantly enhancing the generation efficiency of ND-FWM signal.

Electronic resonant ND-FWM can be utilized for ultra-sensitive detection of highly absorbing

but non-fluorescent molecular species. In Figure 4.6, we demonstrate a micro-spectroscopy application on neocyanine which has an intense absorption (ε~180,000$M^{-1}cm^{-1}$) but non-detectable fluorescence in solution due to extremely short excited lifetime (~ ps). The shaped incident pulse spectrum is tuned to coincide with the absorption peak of neocyanine around 772nm. Under this resonance condition, the generated ND-FWM signal from neocyanine solution rapidly goes up with increasing concentration. As expected, the spectral shapes remain similar as increasing solute concentration.

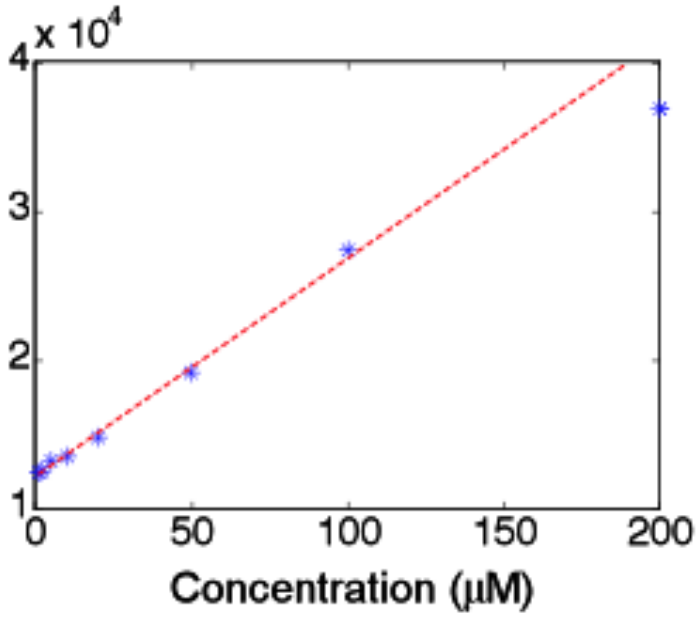


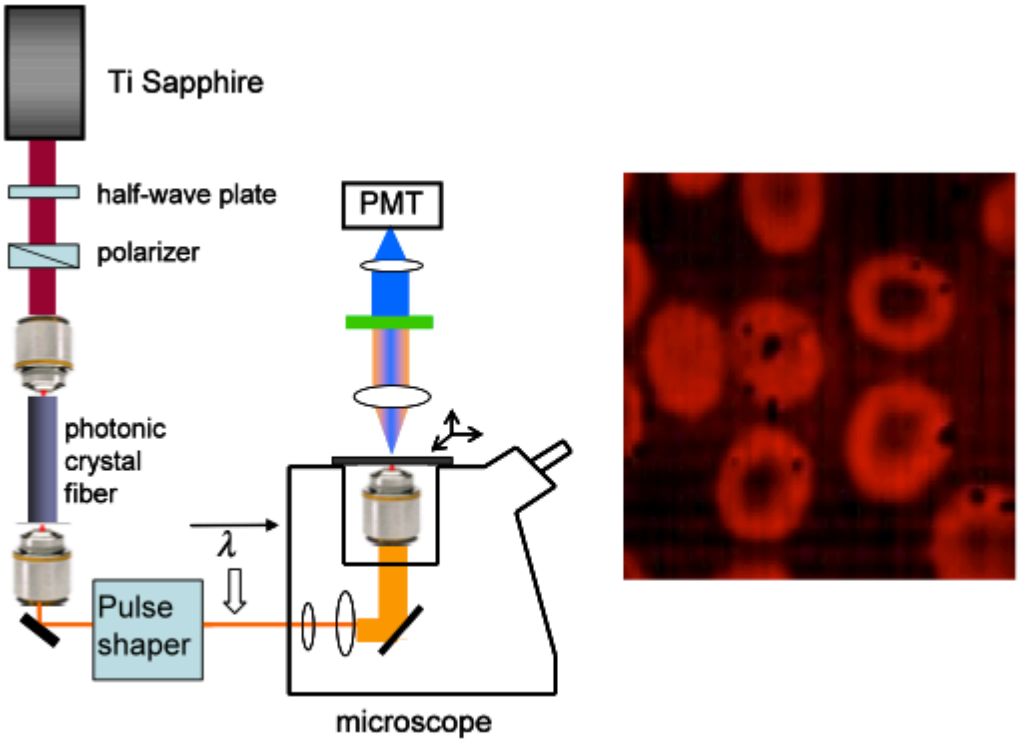
Figure 4.6: Concentration dependence of the ND-FWM signal. The linear dependence at low concentration indicates an optical heterodyne effect. The detection sensitivity is around 1 μM.

The measurement in Figure 4.6 indicates a detection sensitivity of about 1 μM in 100ms. This corresponds to ~ 50 molecules within the laser focus (whose volume is estimated to be about 10^{-16} liter). This represents the highest detection sensitivity of coherent microscopy to the best of our knowledge, about 5 times better than the recently developed triple-resonance CARS micro-spectroscopy (Min et al., 2009b) due to the stronger resonance condition for emission/signal frequency in ND-FWM. We found that optical heterodyne detection also contributes to such a superb sensitivity, as suggested by the linear concentration dependence. This observed linear concentration dependence at the low-concentration limit is due to the interference between the relatively small resonant signal from the low-concentration solute and the strong non-resonant contribution from the ethanol solvent.

Finally, we demonstrate electronic resonant ND-FWM microscopy by imaging non-fluorescent hemoglobin in red blood cells. To be resonant with the *Q* band (~550nm) of hemoglobin absorption spectrum, a visible super-continuum is generated from a photonic crystal fiber with double zero dispersion wavelengths (Murugkar et al., 2007) (Figure 4.7). Such a super-continuum is a promising source for nonlinear imaging of chromophores in the visible spectrum. To generate the super-continuum excitation source, *a* photonic crystal fiber (NL-1.4-775-945, Crystal Fibre A/S, Denmark) with two closely lying zero dispersion

wavelength is used to generate a visible super-continuum. It is 12.5 cm long and is housed inside a hermetically sealed package (Femtowhite 800, Crystal Fibre A/S) that improves ease of light coupling and enhances long term power stability. The pump light from Ti:Sapphire oscillator (Micra 5, Coherent Inc.) is coupled into the fiber by a objective lens (N.A. 0.4) that is mounted on a stable differential 3D translational stage. Average input power of 350mW at 810 nm was typically used to excite the fiber. Averaged power of 0.5~2 mW is used at the sample. The dispersion in the microscope objective is pre-compensated by the prism pair compressor in order to launch the shortest possible (i.e. highest intensity) pulse into the fiber. A home built prism/lens 4-f pulse shaper is used to create the spectral "edge".

Electronic resonant ND-FWM image clearly shows the donut shape of red blood cells. We note that our image exhibits much stronger contrast compared with the reported THG images of red blood cells (Clay et al., 2006), which could be due to the one-photon electronic resonance employed here.



Figure

Figure 4.7: Imaging hemoglobin by electronic resonant ND-FWM microscopy. Experimental apparatus for super-continuum generation from a photonic crystal fiber. The shaped super-continuum excitation pulse is in resonant with the *Q* band of hemoglobin (Hb) and oxy-hemoglobin (HbO_2) absorption spectra. The spectrum of ND-FWM signal is illustrated in green/blue (<570nm). Shown here are red blood cells prepared from fresh blood drawn from a healthy volunteer. The donut shape of red blood cells is clearly illustrated. Scale bar: 10 μm.

As a single-beam multiphoton imaging modality, ND-FWM microscopy provides a distinct nonlinear contrast mechanism based on bulk-sensitive electronic polarization. This technique not only allows 3D high-resolution imaging of live cells and tissues without labeling, but also

enables sensitive detection and mapping of biologically important chromophores such as hemoglobin, beta-carotene, cytochrome *c* and rhodopsin.

References:

Armani, A.M., Kulkarni, R.P., Fraser, S.E., Flagan, R.C., and Vahala, K.J. (2007). Label-Free, Single-Molecule Detection with Optical Microcavities. Science *317*, 783–787.

Barad, Y., Eisenberg, H., Horowitz, M., and Silberberg, Y. (2003). Nonlinear scanning laser microscopy by third harmonic generation. SPIE MILESTONE SERIES MS *175*, 47–49.

Boyd, R.W. (2008). Nonlinear Optics, Third Edition (Academic Press).

Boyer, D., Tamarat, P., Maali, A., Lounis, B., and Orrit, M. (2002). Photothermal Imaging of Nanometer-Sized Metal Particles Among Scatterers. Science *297*, 1160–1163.

Cheng, J.-X., and Xie, X.S. (2002). Green's function formulation for third-harmonic generation microscopy. J. Opt. Soc. Am. B *19*, 1604–1610.

Cheng, J.-X., and Xie, X.S. (2004). Coherent Anti-Stokes Raman Scattering Microscopy: Instrumentation, Theory, and Applications. J. Phys. Chem. B *108*, 827–840.

Clay, G.O., Millard, A.C., Schaffer, C.B., Aus-der-Au, J., Tsai, P.S., Squier, J.A., and Kleinfeld, D. (2006). Spectroscopy of third-harmonic generation: evidence for resonances in model compounds and ligated hemoglobin. J. Opt. Soc. Am. B *23*, 932–950.

Débarre, D., Supatto, W., Pena, A.-M., Fabre, A., Tordjmann, T., Combettes, L., Schanne-Klein, M.-C., and Beaurepaire, E. (2006). Imaging lipid bodies in cells and tissues using third-harmonic generation microscopy. Nature Methods *3*, 47–53.

Denk, W., Strickler, J.H., and Webb, W.W. (1990). Two-photon laser scanning fluorescence microscopy. Science *248*, 73–76.

Dudovich, N., Oron, D., Silberberg, Y., and others (2002). Single-pulse coherently controlled nonlinear Raman spectroscopy and microscopy. Nature *418*, 512–514.

Evans, C.L., and Xie, X.S. (2008). Coherent anti-Stokes Raman scattering microscopy: chemical imaging for biology and medicine. Annu. Rev. Anal. Chem. *1*, 883–909.

Fu, D., Ye, T., Matthews, T.E., Chen, B.J., Yurtserver, G., and Warren, W.S. (2007). High-resolution in vivo imaging of blood vessels without labeling. Opt. Lett. *32*, 2641–2643.

Gannaway, J.N., and Sheppard, C.J.R. (1978). Second-harmonic imaging in the scanning optical microscope. Optical and Quantum Electronics *10*, 435–439.

Hellwarth, R., and Christensen, P. (1975). Nonlinear Optical Microscope Using Second Harmonic Generation. Appl. Opt. *14*, 247–248.

Hiki, S., Mawatari, K., Hibara, A., Tokeshi, M., and Kitamori, T. (2006). UV Excitation Thermal Lens Microscope for Sensitive and Nonlabeled Detection of Nonfluorescent Molecules. Anal. Chem. *78*, 2859–2863.

Ignatovich, F.V., and Novotny, L. (2006). Real-Time and Background-Free Detection of Nanoscale Particles. Phys. Rev. Lett. *96*, 013901.

Kneipp, K., Wang, Y., Kneipp, H., Perelman, L.T., Itzkan, I., Dasari, R.R., and Feld, M.S. (1997). Single Molecule Detection Using Surface-Enhanced Raman Scattering (SERS). Phys. Rev. Lett. *78*, 1667–1670.

Lasne, D., Blab, G.A., De Giorgi, F., Ichas, F., Lounis, B., and Cognet, L. (2007). Label-free optical imaging of mitochondria in live cells. Optics Express *15*, 14184–14193.

Li, H., Harris, D.A., Xu, B., Wrzesinski, P.J., Lozovoy, V.V., Dantus, M., and others (2008). Coherent mode-selective Raman excitation towards standoff detection. Opt. Express *16*, 5499–5504.

Lu, S., Min, W., Chong, S., Holtom, G.R., and Xie, X.S. (2010). Label-free imaging of heme proteins with two-photon excited photothermal lens microscopy. Applied Physics Letters *96*, 113701–113701–3.

Min, W., Lu, S., Chong, S., Roy, R., Holtom, G.R., and Xie, X.S. (2009a). Imaging chromophores with undetectable fluorescence by stimulated emission microscopy. Nature *461*, 1105–1109.

Min, W., Lu, S., Holtom, G.R., and Xie, X.S. (2009b). Triple-Resonance Coherent Anti-Stokes Raman Scattering Microspectroscopy. ChemPhysChem *10*, 344–347.

Min, W., Lu, S., Rueckel, M., Holtom, G.R., and Xie, X.S. (2009c). Near-Degenerate Four-Wave-Mixing Microscopy. Nano Lett. *9*, 2423–2426.

Murugkar, S., Brideau, C., Ridsdale, A., Naji, M., Stys, P.K., and Anis, H. (2007). Coherent anti-Stokes Raman scattering microscopy using photonic crystal fiber with two closely lying zero dispersion wavelengths. Opt. Express *15*, 14028–14037.

Novotny, L., and Hecht, B. (2006). Principles of Nano-Optics (Cambridge University Press).

Pawley, J.B. (2006). Handbook of Biological Confocal Microscopy.

Potma, E.O., de Boeij, W.P., and Wiersma, D.A. (2000). J. Opt. Soc. Am. B *17*, 1678.

Sfeir, M.Y., Wang, F., Huang, L., Chuang, C.-C., Hone, J., O'brien, S.P., Heinz, T.F., and Brus, L.E. (2004). Probing electronic transitions in individual carbon nanotubes by Rayleigh scattering. Science *306*, 1540–1543.

Shen, Y.R. (1988). The Principles of Nonlinear Optics.

Squier, J., Muller, M., Brakenhoff, G., and Wilson, K.R. (1998). Third harmonic generation microscopy. Opt. Express *3*, 315–324.

Weiner, A.M. (2000). Femtosecond pulse shaping using spatial light modulators. Review of Scientific Instruments *71*, 1929–1960.

Raman Microscopy: Developments and Applications Academic Press (1996).

Part II (Chapters 5-9):

Genome Analysis at the Single Cell Level

Summary

With the rapid developments of high throughput sequencing techniques, genome and transcriptome analyses have been routinely done in whole genome with single nucleotide resolution. However, challenges remain. The dynamic changes in DNA molecules generate intra-sample genome heterogeneity. Even with the same genome content, different cells originated form the same cell often have very different transcriptome profile in a functional organism. These information are often masked by performing ensemble analysis on genome and transcriptome.

In this part of the thesis, we first introduce a whole genome amplification method with high genome coverage (~93%) in sequencing human cells (Chapter 6: Whole genome

amplification and sequencing of single human cells). We then use the technique to study meiotic recombinations in single sperm (Chapter 7: Genome-wide study of meiotic recombination in an individual by whole genome sequencing of single sperm cells). We further develop a technique that enables digital counting of genome fragments and whole genome haplotyping in single cells (Chapter 8: Digital whole genome amplification). And we finally introduce our ongoing effort on single cell transcriptome amplification and exploring the genome accessibility at the single cell level (Chapter 9: Single cell transcriptome and genome accessibility studies). Through the development of techniques probing single cell genome, transcriptome and possibly epigenome, we hope to provide a toolbox for studying biological processes with genome-wide and single cell resolution.

Contributions

This part of the thesis involved close collaborations with Dr. Chenghang Zong, Alec R. Chapman, Zi He, Jenny Lu at Harvard, and Wei Fan at Peking University.

In Chapter 6: Whole genome amplification and sequencing of single human cells, Dr. Zong, I and Prof. Xie conceived the idea and initially designed the project. Dr. Zong and I performed most of the experiments. Dr. Zong and A.R. Chapman analyzed data. A.R. Chapman participated in daily discussions and helped design the project.

In Chapter 7: Genome-wide study of meiotic recombination in an individual by whole

genome sequencing of single sperm cells, Prof. Tang and Prof. Xie conceived the idea, Prof. Tang, I and Prof. Xie and Dr. Zong designed the experiments. I performed the experiments with the help of Prof. Tang. W. Fan, A. Chapman, Prof. Ruiqiang Li, Mingyu Yang, Jinsen Li, Ping Zhu and I performed the data analysis. Dr. Zong and I contributed to the experimental method (MALBAC) used in this project and participated in discussions.

In Chapter 8: Digital whole genome amplification, I conceived the general idea and discussed with Prof. Tang and Prof. Xie. A. Chapman, Prof. Xie and I designed the experiments. I performed most of the experiments with the help of Z. He and J. Lu. A. Chapman analyzed the data and participated in daily discussions. Dr. Zong and I contributed to the experimental method (MALBAC) used in this project.

In Chapter 9: Single cell transcriptome and genome accessibility studies, I and Prof. Xie conceived the idea. Prof. Xie, Z.He, I and A. Chapman designed the transcriptome experiments. Prof. Xie, I and J. Lu designed the genome accessibility experiments. Z. He and I performed the transcriptome experiments. J. Lu and I performed the genome accessibility experiments. A. Chapman, Z. He and J. Lu performed data analysis. Dr. Zong participated in discussion and contributed to the experimental method (MALBAC) used in this project.

Chapter 5

Single Cell Genomics: An Overview

5.1 Genome Analysis at the Single Cell Level

With the consensus human genome sequenced and assembled, we have entered into the post-genomic era (Lander et al., 2001; Venter et al., 2001), which has enabled an explosion of modern high throughput genotyping technologies (Mckernan et al.; Braslavsky et al., 2003; Margulies et al., 2005; Gresham et al., 2008). More and more individual humans from different genetic backgrounds have been genotyped and widespread genetic variations (about one variant per kilobases) between individuals are found (Consortium, 2010; Project, 2011), which indicates the instable nature of the human genome.

On a much smaller time-scale, the genome of different cells also exhibit noticeable differences in the lifetime of an individual. Somatic mosaicism of genome contents such as copy number variations (CNVs) was found in differentiated human tissues from multiple individuals (Piotrowski et al., 2008). L1 retrotransposon activity was found in human neural progenitor cells which might have caused the CNVs observed in different regions of the aged human brain (Coufal et al., 2009). More surprisingly, widespread genome differences were found in identical twins which suppose to have identical genome (Bruder et al., 2008). Genome instability generates these somatic variations. Although most of time, these variations might not have significant functional effect, every now and then, one or few cells accumulate enough mutations to outgrow the other cells. And this was hypothesized to be the origin of tumor genesis (Hanahan and Weinberg, 2000). During the development of cancer, such instability often retains and it generates intratumoral genetic heterogeneity (Campbell et al., 2010; Yachida et al., 2010). These cells with different genome often response differently to drugs (Friedlander et al., 1984; Szerlip et al., 2012), which makes it difficult to treat cancers. To understand how these genetic heterogeneities are generated, especially when we are trying to analyze the rare variants in a heterogeneous sample, we must first be able to profile the genome at the single cell level, because the rare variants are often submerged in the ensemble sequencing data because of the limitation on sequencing depth and accuracy (Shendure and Ji, 2008). Single cell genome analysis has been applied successfully to study breast cancer, unveiling the punctuated nature of tumor development in several human breast cancer cases (Navin et al., 2011).

Single cell genome analysis is particularly important with species that are not culturable in labs, such as most microbes. The approach has been successfully used to study TM7, candidate phylum with clinical relevance (Marcy et al., 2007), and marine microbes, etc (Woyke et al., 2009, 2011). Single cell genome sequencing is also necessary when the sample under analysis is rare, such as in the case of using IVF embryos for preimplantation screening (Mastenbroek et al., 2007; Harper et al., 2008) and circulating tumor cells (Cristofanilli et al., 2005; Paterlini-Brechot and Benali, 2007).

To analyze genome at the single cell level, we need a reliable method to amplify the whole genome from a single cell to an enough amount for downstream genetic analysis. Several whole genome amplification methods have been reported, such as PEP-PCR (Snabes et al., 1994; Dietmaier et al., 1999), DOP-PCR (Kuukasjärvi et al., 1997) and Multiple Displacement Amplification (Dean et al., 2001, 2002). These methods however, are often limited by significant amplification bias and low genome coverage. To comprehensively study genome at the single cell level, an improved whole genome amplification method is needed to evenly amplify the whole genome with high accuracy. That is the focus of Chapter 6, in which we introduce a method MALBAC with significantly improved coverage and evenness throughout the genome.

Gametes are haploid cells that fuse with another cell during fertilization in organisms

undergoing sexual reproduction. Each gamete cell is very different in genome content compared to other gametes because of the recombination activity during gametogenesis (Alberts et al., 2007). In Chapter 7, using MALBAC, we examine the genome from each of the 99 sperm, by which we study the genome-wide distribution of an individual with high resolution.

A challenge for genome analysis for diploid organisms is the genome phase problem (Bansal et al., 2011; Tewhey et al., 2011). In a cell of most higher organisms, two copies of homologue chromosomes coexist. The association of different genetic features on a chromosome, known as the phase information, is important for the correct interpretation of the genome, which complicates the genome analysis. This is particularly challenging for single cell genome analysis. In Chapter 8, we introduce a method named digital whole genome amplification to solve this problem without having to isolate each chromosome using complicated devices.

5.2 Transcriptome analysis at the single cell level

A fundamental question in biology is how a single embryo develops into a complete organism consisting of many types of cells with very different function (Gilbert, 2010). Although having identical or very similar genome, a developing embryo or tissue often consists of many different types of cells that are spatially organized into functional structures (Calvi et al.,

2003; Barker et al., 2007). The cell types that are responsible for renewing the tissue or are essential for the structures often consist of a very small fraction of the entire tissue (Toma et al., 2001; Wagers and Weissman, 2004). In this situation, it is crucial to isolate individual cells from a tissue with complex microenvironment and study the transcriptome of each single cell individually.

It is also becoming clear that the transcription is rather noisy for a significant portion of genes, exhibiting 'bursting' transcription dynamics (Taniguchi et al., 2010; Suter et al., 2011), which may be a result of the intrinsic stochasticity because of the low DNA copy number (Raj and van Oudenaarden, 2008) or a result of the changing microenvironment (Thattai and Oudenaarden, 2004). With such widespread stochasticity, it is surprising how cells maintain their functions in a given condition. Biochemical networks maintaining functional robustness in cells have been found (Shinar and Feinberg, 2010), and studying the transcriptome at the single level can contribute to testing these networks, and finding new players, and even finding new networks conferring robustness of cellular functions (Stelling et al., 2004; Kærn et al., 2005).

Polymerase chain reactions (PCR) was invented in the 1980s (Saiki et al., 1988) and had quickly revolutionized molecular biology since then (Sambrook and Russell, 2001). PCR has been used to amplify DNA and RNA molecules from single cells for genetic testing (Handyside et al., 1992; Wells et al., 2002) and gene expression studies (Eberwine et al.,

1992; Brand and Perrimon, 1993), which reveal different subpopulations with distinct transcription profile. These studies, however, are often limited by the throughput of the testing loci. With the development of microfluidic chips, it is now possible to measure tens to hundreds of genes simultaneously in a single cell (Wheeler et al., 2003; Dalerba et al., 2011).

With the development of modern high throughput genotyping techniques, thousands or even the complete ~20,000 genes can be measured simultaneously (Mckernan et al.; Braslavsky et al., 2003; Margulies et al., 2005; Gresham et al., 2008). The further development of high throughput transcriptome sequencing provides information that cannot be accessible by other methods (Mortazavi et al., 2008). These genotyping techniques require large amount of starting materials (~1 microgram of RNA, corresponding to ~10^7 mammalian cells). Transcriptome amplification to microgram level of DNA is required for these high throughput genotyping studies on single cells (Kurimoto et al., 2006, 2007; Tang et al., 2009; Islam et al., 2011).

Due to the amplification unevenness and bias from ~100 femtogram of mRNA to the microgram level, inferring the relative gene expression level in a single cell is often tricky and must be carefully designed with controls for these high throughput genotyping experiments (Tang et al., 2010, 2011). In Chapter 9, we introduce a new amplification method for transcriptome analysis from single human cells. We further use this method to study genome and transcriptome of the same single human cells.

Epigenetic factors, such as methylation (Bird, 2002) and histone modifications (Strahl et al., 2000) are fundamental in shaping the transcriptome profile of mammalian cells. However, there is currently no established method for studying these epigenetic factors genome-wide at the single cell level. In Chapter 9, we introduce our initial efforts to fill in this 'missing link' between single cell genome and transcriptome analysis. Using the method we introduce in Chapter 8, we are studying the genome-wide chromosome accessibility of single human cells.

References:

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). Molecular Biology of the Cell (Garland Science).

Bansal, V., Tewhey, R., Topol, E.J., and Schork, N.J. (2011). The next phase in human genetics. Nature Biotechnology *29*, 38–39.

Barker, N., Es, J.H. van, Kuipers, J., Kujala, P., Born, M. van den, Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P.J., et al. (2007). Identification of stem cells in small intestine and colon by marker gene Lgr5. Nature *449*, 1003–1007.

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S.R. (2003). Sequence information can be obtained from single DNA molecules. PNAS *100*, 3960–3964.

Bruder, C.E.G., Piotrowski, A., Gijsbers, A.A.C.J., Andersson, R., Erickson, S., Diaz de Ståhl, T., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A., et al. (2008). Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. The American Journal of Human Genetics *82*, 763–771.

Calvi, L.M., Adams, G.B., Weibrecht, K.W., Weber, J.M., Olson, D.P., Knight, M.C., Martin, R.P., Schipani, E., Divieti, P., Bringhurst, F.R., et al. (2003). Osteoblastic cells regulate the haematopoietic stem cell niche. Nature *425*, 841–846.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.-L., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109–1113.

Consortium, T.1000 G.P. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127–1131.

Cristofanilli, M., Hayes, D.F., Budd, G.T., Ellis, M.J., Stopeck, A., Reuben, J.M., Doyle, G.V., Matera, J., Allard, W.J., Miller, M.C., et al. (2005). Circulating Tumor Cells: A Novel Prognostic Factor for Newly Diagnosed Metastatic Breast Cancer. JCO *23*, 1420–1430.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. PNAS *99*, 5261–5266.

Dean, F.B., Nelson, J.R., Giesler, T.L., and Lasken, R.S. (2001). Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. Genome Res *11*, 1095–1099.

Dietmaier, W., Hartmann, A., Wallinger, S., Heinmöller, E., Kerner, T., Endl, E., Jauch, K.-W., Hofstädter, F., and Rüschoff, J. (1999). Multiple Mutation Analyses in Single Tumor Cells with Improved Whole Genome Amplification. The American Journal of Pathology *154*, 83–95.

Friedlander, M.L., Hedley, D.W., and Taylor, I.W. (1984). Clinical and biological significance of aneuploidy in human tumours. J Clin Pathol *37*, 961–974.

Gilbert, S.F. (2010). Developmental Biology, Ninth Edition (Sinauer Associates, Inc.).

Gresham, D., Dunham, M.J., and Botstein, D. (2008). Comparing whole genomes using DNA microarrays. Nat. Rev. Genet. *9*, 291–302.

Hanahan, D., and Weinberg, R. (2000). The Hallmarks of Cancer. Cell *100*, 57–70.

Harper, J., Sermon, K., Geraedts, J., Vesela, K., Harton, G., Thornhill, A., Pehlivan, T., Fiorentino, F., SenGupta, S., Die-Smulders, C. de, et al. (2008). What next for preimplantation genetic screening? Hum. Reprod. *23*, 478–480.

Kuukasjärvi, T., Tanner, M., Pennanen, S., Karhu, R., Visakorpi, T., and Isola, J. (1997). Optimizing DOP-PCR for universal amplification of small DNA samples in comparative genomic hybridization. Genes Chromosomes Cancer *18*, 94–101.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Marcy, Y., Ouverney, C., Bik, E.M., Losekann, T., Ivanova, N., Martin, H.G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D.A., et al. (2007). Inaugural Article: Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proceedings of the National Academy of Sciences *104*, 11889–11894.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Mastenbroek, S., Twisk, M., van Echten-Arends, J., Sikkema-Raddatz, B., Korevaar, J.C., Verhoeve, H.R., Vogel, N.E.A., Arts, E.G.J.M., De Vries, J.W.A., Bossuyt, P.M., et al. (2007). In vitro fertilization with preimplantation genetic screening. New England Journal of Medicine *357*, 9–17.

MCKERNAN, K., BLANCHARD, A., Kotler, L.E.V., and COSTA, G. Reagents, Methods, and Libraries for Bead-Based Sequencing.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Paterlini-Brechot, P., and Benali, N.L. (2007). Circulating tumor cells (CTC) detection: clinical impact and future directions. Cancer Letters *253*, 180–204.

Piotrowski, A., Bruder, C.E.G., Andersson, R., de Ståhl, T.D., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A., et al. (2008). Somatic mosaicism for copy number variation in differentiated human tissues. Human Mutation *29*, 1118–1124.

Project, the 1000 G. (2011). Variation in genome-wide mutation rates within and between human families. Nature Genetics *43*, 712–714.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology *26*, 1135–1145.

Snabes, M.C., Chong, S.S., Subramanian, S.B., Kristjansson, K., DiSepio, D., and Hughes, M.R. (1994). Preimplantation single-cell analysis of multiple genetic loci by whole-genome amplification. PNAS *91*, 6181–6185.

Szerlip, N.J., Pedraza, A., Chakravarty, D., Azim, M., McGuire, J., Fang, Y., Ozawa, T., Holland, E.C., Huse, J.T., Jhanwar, S., et al. (2012). Intratumoral heterogeneity of receptor tyrosine kinases EGFR and PDGFRA amplification in glioblastoma defines subpopulations with distinct growth factor response. Proc Natl Acad Sci U S A *109*, 3041–3046.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J., and Schork, N.J. (2011). The importance of phase information for human genomics. Nature Reviews Genetics *12*, 215–223.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The Sequence of the Human Genome. Science *291*, 1304–1351.

Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., Malmstrom, R., Stepanauskas, R., and Cheng, J.-F. (2011). Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. PLoS ONE *6*, e26161.

Woyke, T., Xie, G., Copeland, A., González, J.M., Han, C., Kiss, H., Saw, J.H., Senin, P., Yang, C., Chatterji, S., et al. (2009). Assembling the Marine Metagenome, One Cell at a Time. PLoS ONE *4*, e5299.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Chapter 6

Whole Genome Amplification and Sequencing of Single Human Cells

6.1 Summary and Background

As the basic unit of life, a single cell carries the blueprint of life that consists of individual DNA molecules of specific sequences (Krebs et al., 2009). Genomic instability results in variation of DNA molecules originating from the same lineage, which underlies important biological processes such as evolution and tumor genesis (Futuyma, 2009; Hanahan and Weinberg, 2011) . However, such genetic variations are often masked by ensemble genetic analysis such as whole genome sequencing using large amount of starting materials. Identifying these genomic differences among cells and studying genome comprehensively with very limited starting materials (such as a single cell) are fundamental to many biological investigations and are important for many medical applications.

Single-cell whole-genome amplification methods have been reported but are hindered by amplification bias, resulting in low genome coverage. In this chapter, we introduce a new amplification method: Multiple Annealing and Looping Based Amplification Cycles (MALBAC) that achieves ~93% genome coverage ≥ 1x for a single human cell at ~30x mean sequencing depth. We demonstrate probing digitized copy number variations as well as detection of single nucleotide variations (SNVs) with an overall ~76% efficiency for a single human cancer cell. By sequencing three descendent cells from a single cell, we call SNVs with a low false positive rate similar to that of bulk sequencing, which allows us to directly measure the genome-wide mutation rate in the cancer cell line.

Single molecule and single cell studies have been routinely revealing individual behaviors that are otherwise hidden in bulk measurements (Elowitz et al., 2002; Li and Xie, 2011). In a human cell, the genetic information is encoded in 46 single DNA molecules—chromosomes. The variations occurring in these single molecules, such as single nucleotide variations (SNVs) and copy number variations (CNVs) (Negrini et al., 2010), are the driving forces in important biological processes such as evolution and tumor genesis (Futuyma, 2009; Hanahan and Weinberg, 2011). Such dynamic variations are reflected in the genomic heterogeneity among a population of cells, which demands characterization of genomes at the single cell level. This is particularly the case for cancer because of the well-established intratumoral genetic heterogeneity among cells (Lengauer et al., 1998; Campbell et al., 2010; Yachida et

al., 2010). Single cell genomics analysis is also particularly necessary when the number of cells available is limited to few or one, such as prenatal testing samples (Lo et al., 2010; Kitzman et al., 2012), circulating tumor cells (Nagrath et al., 2007), and forensic specimens (Hanson and Ballantyne, 2005).

Prompted by rapid developments of the next generation sequencing techniques (Metzker, 2010), there have been several reports of single cell whole genome sequencing (Fan et al., 2011; Navin et al., 2011; Wang et al., 2012). The prevailing method is using whole genome amplification (WGA) before sequencing (Telenius et al., 1992; Zhang et al., 1992, 2006; Dean et al., 2002; Lao et al., 2008). However, due to the amplification unevenness and bias, single-cell sequencing is hindered by low genome coverage. Polymerase chain reaction (PCR) has been a gold standard for DNA amplification of specific sequences (Saiki et al., 1988). PCR-based WGA relies on exponential amplification with random primers, which introduces strong sequence-dependent bias. Multiple Displacement Amplification (MDA) with phi29 DNA polymerase has provided improvements over the PCR-based methods, but still exhibits considerable bias due to nonlinear amplification (Dean et al., 2002).

6.2 Multiple Annealing and Looping Based Amplification (MALBAC)

We have developed a new WGA method named Multiple Annealing and Looping Based Amplification Cycles (MALBAC), which introduces close-to-linear preamplification to reduce the bias pertinent to nonlinear amplification. Picograms of DNA fragments (~10 to

100kb) from a single human cell serve as templates for amplification with MALBAC (Figure 6.1).

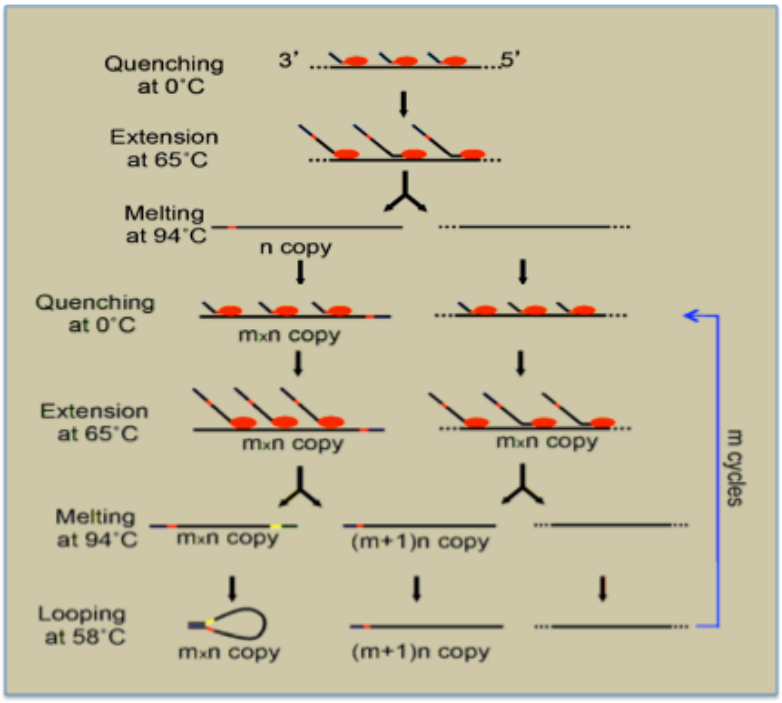


Figure 6.1: Multiple Amplification and Looping-based Amplification Cycles (MALBAC). The DNA templates are dehybridized onto single-stranded DNA molecules at 94°C, MALBAC primers anneal randomly to single-stranded DNA molecules at 0°C and are extended by a polymerase with displacement activity at elevated temperatures, creating semi-amplicons. In the following m temperature cycles, single stranded amplicons and the genomic DNA are used as template to produce full amplicons and additional semi-amplicons, respectively. For full amplicons, the 3' end is complementary to the sequence on the 5' end. The two ends hybridize will form the looped DNA, which can efficiently prevents the full amplicon from being used as template, therefore warrant a linear mechanism of amplification. After the five cycles of linear preamplification, only the full amplicons can be exponentially amplified by PCR, generating microgram level of DNA material for sequencing experiments.

The amplification is initiated with a pool of random primers, each having a fixed anchor sequence at the 5' end and eight degenerate nucleotides at the 3' end, that evenly hybridize to the templates at 0^{o}C. At elevated temperature 65^{o}C, DNA polymerases with strand displacement activity are used to generate semi-amplicons with variable lengths, which are then melt off from the template at 94^{o}C. In the following five cycles of pre-amplification, the full amplicons, each with the complimentary ends, allow the formation of DNA loops to prevent the amplicons from further amplification as well as self and cross hybridization. The linear MALBAC pre-amplification is followed by exponential amplification of the full amplicons by PCR in order to generate micrograms of DNA required for next generation sequencing (Figure 6.2).

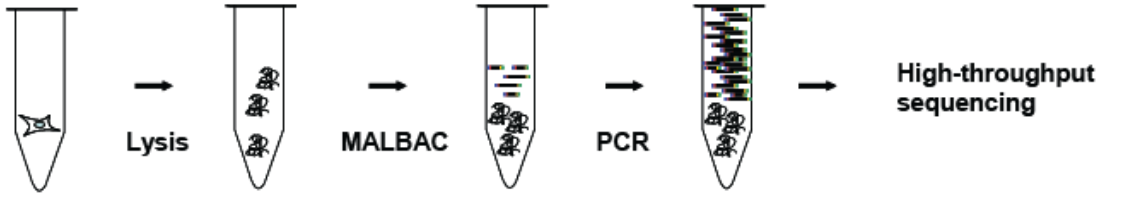


Figure 6.2: High throughput sequencing of a single cell. Lysis of a single cell is followed by dehybridization of genomic DNA molecules into single-stranded DNAs. MALBAC pre-amplification prior to additional PCR amplification is performed before high-throughput sequencing.

In detail, we obtained the SW480 colorectal adenocarcinoma cell line from American Type Culture Collection (ATCC, Rockville). SW480 cells are maintained in ATCC-formulated Leibovitz's L-15 Medium supplemented with 10% fetal bovine serum (ATCC), 100 I.U./ml

Penicillin and 100 mg/ml Streptomycin (ATCC). The cells are treated with 0.25% Trypsin-EDTA, followed by washing and dilution in PBS. Single cells are then mouth pipetted into individual PCR tubes. After briefly spinning down the single cells to the bottom of PCR tubes, 5µl of freshly prepared cell lysis buffer containing QIAGEN protease is prepared according to manufacturer's specifications and added into each tube. The lysis of the single cell is performed by the following temperature steps: 50°C 3 hours, 75°C 20 minutes, 80°C 5 minutes. Isolation of single cells can also be performed with other techniques such as laser dissection, microfluidic devices, or flow cytometry. Avoiding DNA contamination from environment and operators is critical for single cell SNV analysis.

The amplification products have a size distribution of ~500bp to 1500bp, and are then used for preparing sequencing libraries for Illumina and SOLiD sequencing platforms. The DNA product from MALBAC amplification can be directly used in constructing the sequencing library for both Illumina and SOLiD with standard procedures. For Illumina sequencing, ~3ug DNA material are provided to a vendor for standard library preparation and sequencing. For SOLiD sequencing, we performed the library preparation and sequencing following the standard protocol of SOLiD 4 system. 3ug DNA material is used as the starting material. We used EZ bead system for emulsion PCR and enrichment. The platform for Illumina sequencing is Hiseq-2000.

6.3 Performance Characterization of MALBAC

Before we used high throughput sequencing to fully characterize the amplification properties. We first used qPCR to check on 11 unique human genome loci. Shown in Table 6.1 are the Ct numbers for a typical MALBAC amplification from a single cell, compared with amplification using ~500pg as a positive control. 14 out of the 16 loci were evenly amplified, indicating the amplification is even at most of the genome positions. We therefore decided to sequence the whole genome of these cells.

qPCR	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr9
Single Cell	22.5	24.4	36.5	24.8	25.4	26.9	25.5	25.2
Positive Contro	22.2	24.5	29.6	24.4	24.4	24.4	25.5	25.1
qPCR	Chr12	Chr13	Chr13	Chr15	Chr16	Chr17	Chr18	Chr19
Single Cell	25.8	23.9	39.0	24.7	21.3	24.3	24.2	27.0
Positive Control	26.4	26.3	30.0	23.8	20.0	25.8	23.7	22.5

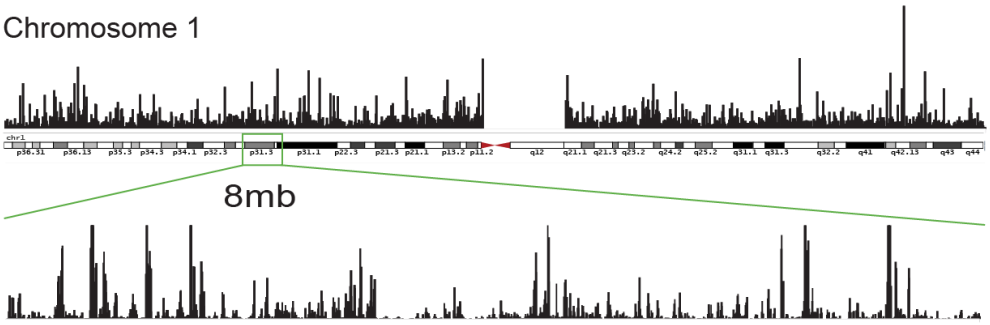
Table 6.1: quantitative PCR result of a typical single cell whole genome amplification reaction using MALBAC. Shown here are the Ct numbers of randomly selected 16 loci each on a different chromosome. The single cell results are consistent with the positive control containing 500pg of DNA as starting materials. 14 out of the 16 loci are amplified evenly. The qPCR result is consistent with the ~90% genome coverage with 30x sequencing depth for single cells. Ct numbers of negative controls are all larger than 30 cycles for the above q-PCR primer pairs.

With ~25x mean sequencing depth, we consistently achieved ~85% of genome coverage at ≥1x depth and ~93% coverage with 30x mean sequencing depth. As a comparison, we

performed MDA on a single cell from the same cancer cell line. At 25x mean sequencing depth, MDA covered 72% of the genome at ≥1x coverage. While significant variations of the coverage have been reported for MDA (Zhang et al., 2006; Hou et al., 2012; Wang et al., 2012; Xu et al., 2012), The MALBAC coverage is consistently reproducible.

The higher amplification coverage of MALBAC is reflected in the evenness of amplification shown in Figure 6.3, in which the amplification coverage map is plotted for both MALBAC and MDA on a single cell. While the MDA exhibits 'spiky' amplification showing local over and under amplification of random regions, the single cell amplified with MALBAC is even at a global scale, and exhibits much less local amplification variations as well.

MDA:



MALBAC:

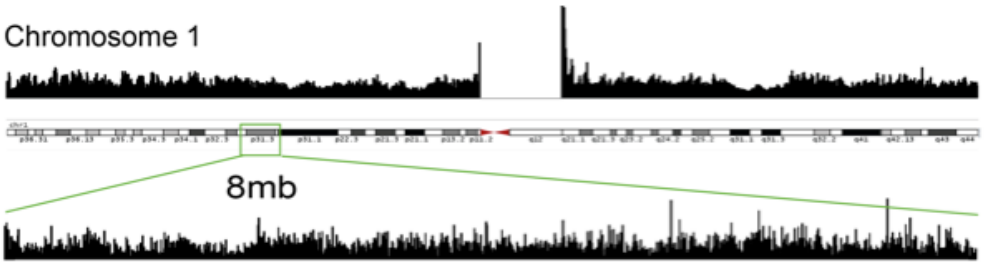


Figure 6.3: The histogram of reads (chromosome 1) of MALBAC compared with MDA amplification of a single SW480 cell. The average sequencing depth is ~25x for both methods.

We used Lorenz curves to evaluate coverage uniformity along the genome. Here, we plotted the cumulative fraction of the total reads possessed by a given cumulative fraction of genome (Figure 6.4). The diagonal line indicates a perfectly uniform distribution of reads, and deviation from the diagonal line indicates an uneven distribution of reads. We compared the Lorenz curves for bulk sequencing, MALBAC, and MDA at ~10x mean sequencing depth. It is evident that MALBAC outperforms MDA in uniformly covering the genome.

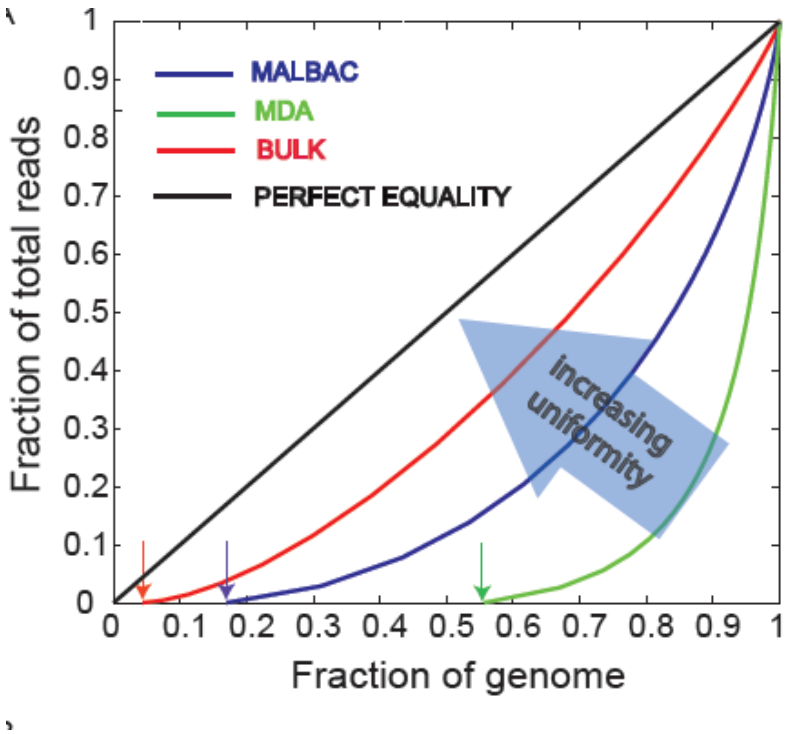


Figure 6.4: Lorenz curves of MALBAC, MDA and bulk sample. A Lorenz curve gives the cumulated fraction of reads as a function of the cumulated fraction of genome. Perfectly uniform coverage would result in a diagonal line and a large deviation from the diagonal is indicative of a biased coverage. All samples are down sampled at 10x depth. The arrows showed the percentage of genome uncovered at 10x sequencing depth.

We also plotted the power spectrum of read density variations to show the spatial scale at which the variations take place, which is important for copy number variation analysis of single cells. Figure 6.5 shows that MALBAC has a power spectrum similar to that of the unamplified bulk. In contrast, the power spectrum of MDA has significantly higher variations in the range of tens of kilobases to tens of megabases.

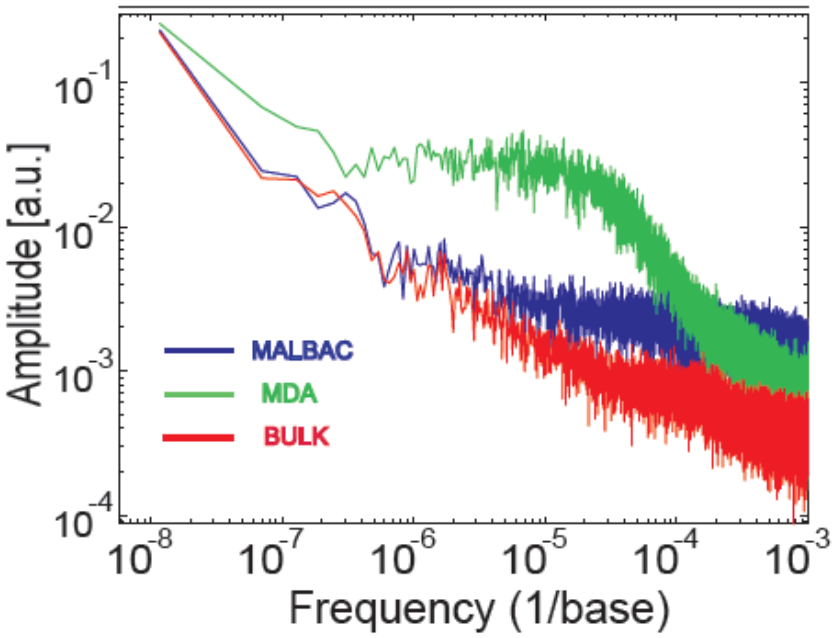


Figure 6.5: Power spectrum of read density throughout the genome (as a function of spatial frequency). MALBAC performs similarly to bulk, while the MDA spectrum contains a large amount of low-frequency components, demonstrating that regions of several megabases suffer from under- and over- amplification. This observation is consistent with the histogram of read depth shown in Figure 6.3.

6.4 Detection of Copy Number Variations (CNVs) in Single Cells

Copy number variation (CNV) is a major form of genome variations and is commonly found in different human individuals (Sebat et al., 2004; Redon et al., 2006). At the single cells level, CNVs are due to insertions, deletions, or multiplications of genome segments, which are frequently observed in almost all categories of human tumors (Beroukhim et al., 2010; Navin et al., 2011; Stephens et al., 2011). MALBAC's immunity to such large-scale bias

makes it amenable to probe CNVs in single cells. We determined the digitized CNVs across the whole genomes of individual cells from the SW480 cancer cell line (Figure 6.6). There are distinct CNV differences among the three individual cancer cells as well as the bulk result, which are difficult to resolve by MDA.

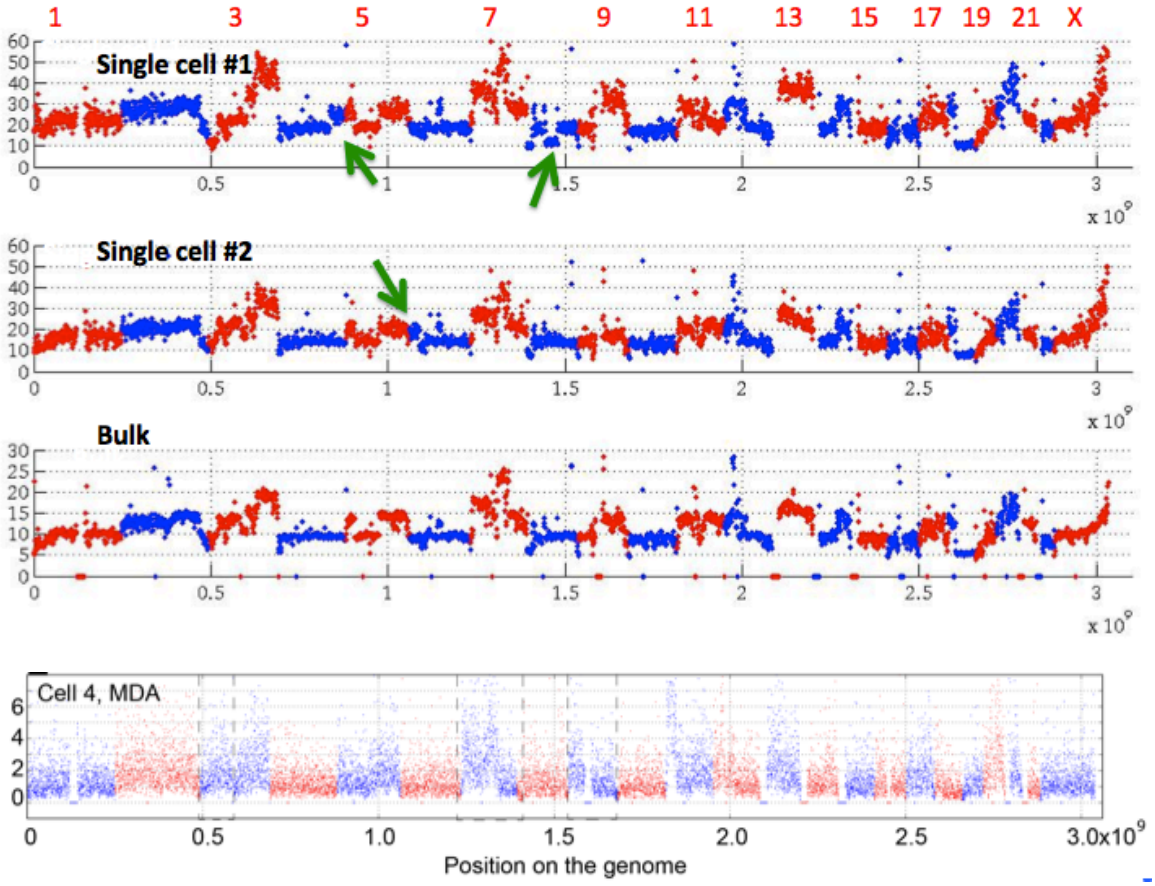


Figure 6.6: CNVs of single cancer cells. Digitized copy numbers across the genome are plotted for two single cells as well as the bulk sample from the SW480 cancer cell line. The bottom panel shows the result based on MDA amplification. The single cells are sequenced at only 0.8x depth, while the bulk and MDA are done at 25x.

For the MALBAC data, we used a hidden Markov model to quantify the CNVs. We confirmed the gross features of CNVs detected by MALBAC using spectral karyotyping (Rochette et al., 2005). For example, both MALBAC-based quantification of CNVs and spectral karyotyping show that chromosome 18 has only one copy while chromosome 17 has three copies in the SW480 cancer cell line. Although the majority of copy numbers are consistent between single cells, we also observe cell-to-cell variations as indicated by the green arrows in Figure 6.6.

6.5 Detection of Single Nucleotide Variations (SNVs) in Single Cells

There were several attempt recently to call SNVs from a single cell by whole genome amplification using MDA (Hou et al., 2012; Wang et al., 2012; Xu et al., 2012). The first challenge in accurate SNV calling from a single cell is substantial human contamination from the environment and the operators, given picograms of DNA from a single human cell. The second challenge is the low detection yield (high false negative), particularly allele dropouts due to amplification bias. The third challenge is false positives associated with amplification and sequencing errors, either by random or systematic (MacArthur, 2012).

In meeting with the first challenge, we took special precautions to decontaminate the reagents and PCR tubes with UV radiation (Woyke et al., 2011) before each experiment was conducted in a restricted clean room. An alternative approach to reduce contamination is by using

microfluidics to isolate cells and perform whole genome amplification (Blainey and Quake, 2011).

In response to the second challenge, MALBAC allowed us to call 2.4 x 10^6 single cell SNVs out of 2.8 x 10^6 detected SNVs in bulk, yielding an 82% detection efficiency, compared to the 41% with MDA (Table 6.2).

	Heterozygous SNVs	Homozygous SNVs	Total SNVs
Bulk			
SNVs	911,958	1,930,204	2,842,162
Single cell MDA			
SNVs Detection efficiency	93,140 (2,828)* 10%	1,238,286 (1,973) 63%	1,331,426 (4,801) 41%
Single cell MALBAC			
SNVs Detection efficiency	756,812 (108,481) 71%	1,539,326 (6,821) 80%	2,296,138 (115,302) 76%

Table 6.2: Comparison of Single cell SNVs and allele dropout for bulk, MDA and MALBAC. The number in the bracket indicates the number of false positives.

In Table 6.2, we listed the heterozygous and homozygous SNVs separately. We then calculated the allele dropout rate for MALBAC compared with MDA. The 7,288 SNVs genotyped as homozygous mutations by MALBAC are actually heterozygous in bulk, which corresponds to a ~1% allele dropout rate in MALBAC. In contrast, with MDA we found 172,563 incorrect homozygous calls, corresponding to an allele dropout rate of ~65%.

The estimation of allele dropout rate is as follows. Here we denote the allele dropout rate as α and N as the number of the heterozygous SNV positions that have enough reads covered for SNV analysis. The number N is related to the coverage of amplification methods. For the sequencing data with MDA, we found 172,565 SNVs that are called homozygous SNVs based on single cell data, but are heterozygous SNVs based on bulk data. This indicates allele dropout of the reference allele. Here we assume the cell is diploid for simplification, and then we have $N\alpha(1-\alpha) = 172{,}565$. In MDA, we also called 93,140 SNVs as heterozygous, which follows $N(1-\alpha)(1-\alpha) = 93{,}140$. With the two above equation, we estimate the allele dropout a of MDA is ~65%. Similarly, we estimated the allele dropout rate for MALBAC is ~1%. This efficiency of amplifying both alleles is contributed by the multiple annealing and amplification cycles in MALBAC.

We now discuss the false positive rate of SNVs. Our MALBAC data contains 1.1×10^5 false positives (Table 6.2), which corresponds to ~10^{-4} false positive rate. Although improving polymerase's error rate is possible, it is not likely such rate can be lower than ~10^{-6}. It is

therefore necessary to have two or three replicates to reduce the false positive. We obtained these replicates by sequencing two or three descendent cells derived from the same cell. The simultaneous appearance of an SNV in the descendent cells would indicate a true SNV. The false positive rate due to random errors can be reduced to $\sim 10^{-8}$ with two descendent cells and $\sim 10^{-12}$ with three descendent cells. The false positives due to systematic sequencing and amplification errors were inspected by comparing two unrelated single cells that are not from the same lineage. After removing both types of errors, we estimated the number of false positives to be less than one per genome with three descendent cells (Table 6.3).

	Heterozygous SNVs	Homozygous SNVs	Total SNVs
Two descendent cells			
SNVs	615,387	1,322,555	1,937,942
Detection efficiency	67%	68%	68%
Newly acquired SNVs	145 (~110)*	3 (0)	148 (~110)
Three descendent cells			
SNVs	660,246	1,577,798	2,238,044
Detection efficiency	72%	81%	80%
Newly acquired SNVs	30 (~0)	5 (0)	35 (~0)

Table 6.3: MALBAC calling of total SNVs and newly acquired SNVs using two and three descendent cells. The number in the bracket indicates the number of false positives.

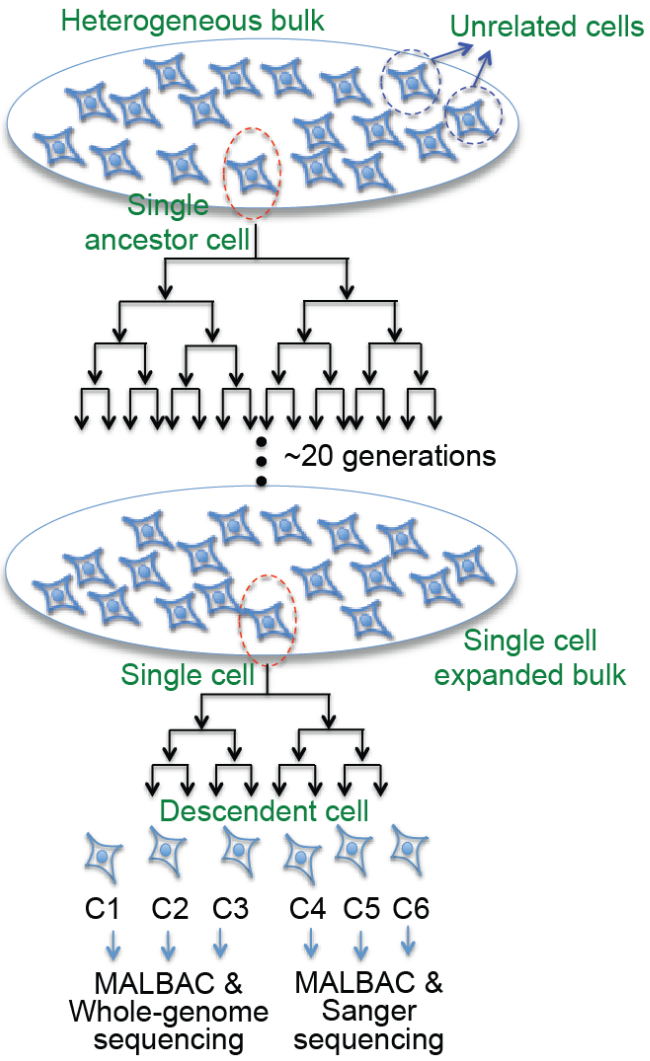


Figure 6.7: Experiment design for calling newly acquired SNVs and estimation of mutation rate of a cancer cell line SW480. A single ancestor cell is chosen and cultured for ~20 generations. The vast majority of cells are used to extract DNA for bulk sequencing to represent the ancestor cell's genome. A single cell from this culture is chosen for another expansion of four generations. The descendent cells are isolated for single cell whole genome amplification. Single cell sample C1, C2, and C3 are used for high-throughput sequencing. Sample C4, C5, and C6 are used for varying SNVs with Sanger sequencing.

With the goal of detecting the genomic difference between cells, we designed an experiment shown in Figure 6.7. First, we clonally expanded a single ancestor cell picked from a heterogeneous population of the SW480 cancer cell line for 20 generations. Then we extracted DNA from this single cell clonal expansion for bulk sequencing, which reflects the genome of the ancestor cell. We then picked a single cell from this clone with the goal of detecting the unique SNVs newly acquired by this cell during the expansion. We grew the cells for another four generations to obtain the descendent cells denoted C1 to C16. We individually sequenced three descendent cells, C1, C2, and C3 after MALBAC amplification, which allows us to remove random amplification errors. Figure 6-8 shows the screening of amplification errors in the case of a pair of descendent cells which resulted in 110 unique SNV candidates (red dots not on the xy axes). The random amplification errors can be further reduced to less than one per genome with improved fidelity of the polymerases. Instead we show that random amplification errors can be eliminated with an additional cell.

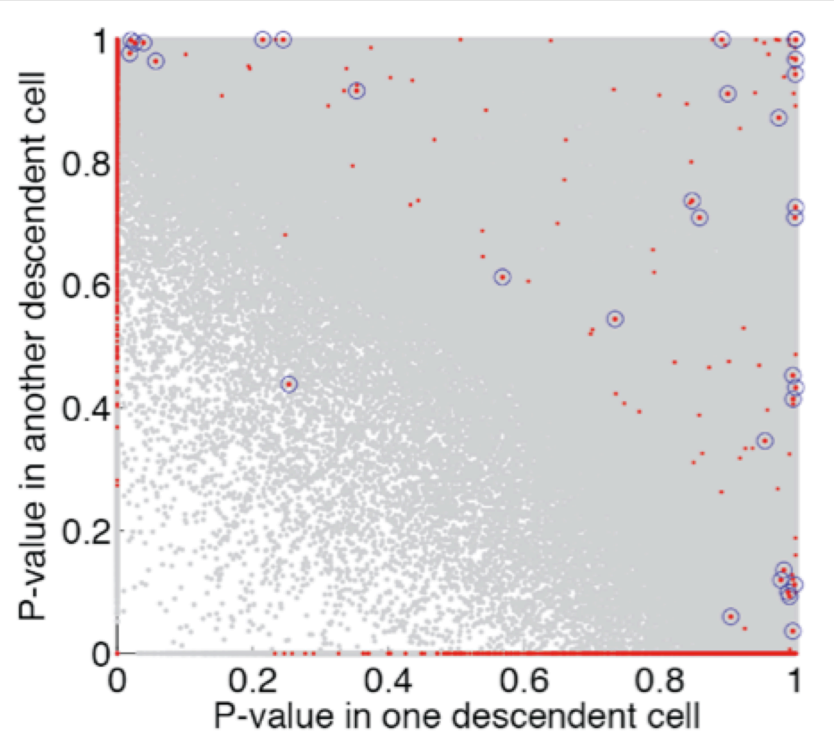


Figure 6.8: The 2D p-value plot of a one-sided binomial test for SNV candidates summed from all pairs of the three descent cells. The cumulated probability of the number of reads of the mutant allele and the total number of reads covering each SNV position is calculated. The grey dots are the SNVs present in the bulk data; the red dots SNVs not present in the bulk. The circled red dots are the 35 newly acquired SNVs during the 20-generation of clonal expansion. We note that all the homozygous SNVs all locate at (1,1) position.

We further filtered out the systematic errors due to homopolymer and tandem repeat sequences. After applying these filters, we detected 35 unique SNVs (circled red dots in Figure 6.8). Their distribution on the chromosomes is shown in Figure 6.9.

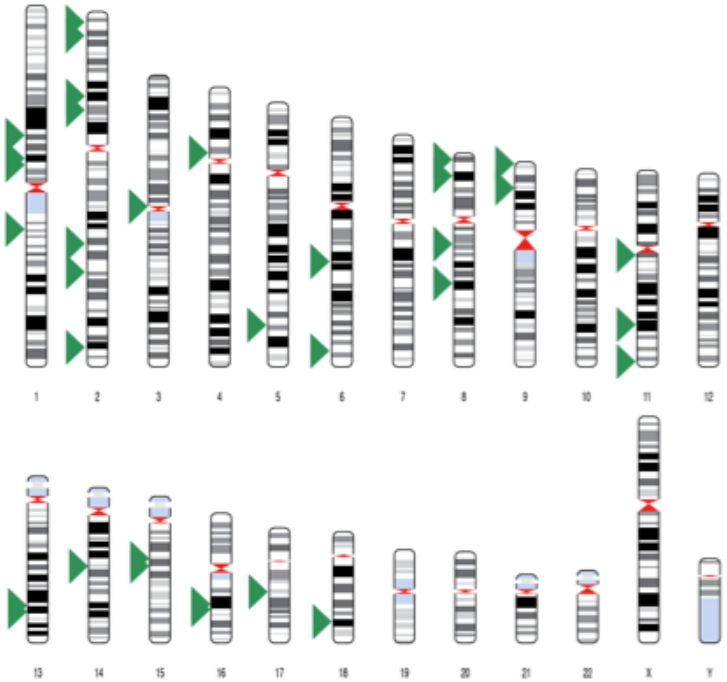


Figure 6.9: The distribution of the 35 newly acquired SNVs on the chromosomes of a single cell

We randomly chose 8 out of total 35 unique SNVs and confirmed that they are not false positives by Sanger sequencing C4-C6, nor false negatives by Sanger sequencing the bulk. As an example, Figure 6.10 shows the MALBAC and Sanger sequencing results of one such SNV.

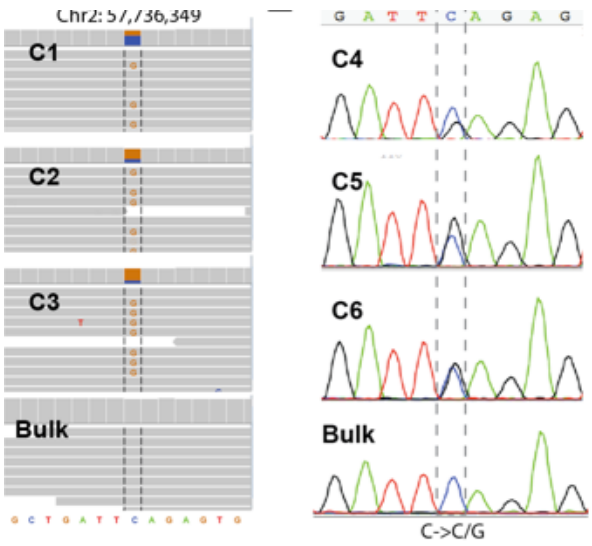


Figure 6.10: Example of a newly acquired SNV. The SNV (C→G) exists in the high throughput data of all three descendent cells but not in the bulk data. The Sanger sequencing of single cells C4, C5, and C6 confirms that this SNV is not a false positive, and the Sanger sequencing of the bulk confirms that this SNV is not a false negative.

The 35 unique SNVs we identified were newly acquired during the 20 cell divisions, which allows us to estimate the mutation rate of the cell line under normal culturing condition. After adjusting for a detection efficiency of 72% for heterozygous SNVs, we estimate that ~49 mutations occurred in the 20 generations, yielding a mutation rate of ~2.5 nucleotides per cell generation. This is the first time a genome-wide mutation rate per cell generation has been directly measured for human somatic cells. Interestingly, the mutation rate of this cancer cell line is compatible with the mutation rates based on germ line studies (Roach et al., 2010; Project, 2011).

Surprisingly, we found that the transition/transversion (Ts/Tv) ratio for the 35 newly acquired SNVs detected is ~0.30, whereas the ratio for the total SNVs of this cell line is 2.01, as expected for common human mutations (Consortium, 2010). To further confirm that this observation is not due to single cell amplification, we sequenced the bulk DNA of the original heterogeneous cultureWe found out that the Ts/Tv for SNVs detected in the single cell expanded bulk but not in original heterogeneous bulk is ~0.75. Both the significantly lower Ts/Tv values indicate that transition is not favored over transversion for newly acquired SNVs in this cancer cell line. While understanding the underlying mechanism of this phenomenon will require similar measurements in other systems, with the ability of precise characterization of CNVs and SNVs, MALBAC can shed light on the individuality, heterogeneity, and dynamics of the genomes of single cells, the basic units of life.

References:

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Blainey, P.C., and Quake, S.R. (2011). Digital MDA for enumeration of total nucleic acid contamination. Nucl. Acids Res. *39*, e19–e19.

Campbell, P.J., Yachida, S., Mudie, L.J., Stephens, P.J., Pleasance, E.D., Stebbings, L.A., Morsberger, L.A., Latimer, C., McLaren, S., Lin, M.-L., et al. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature *467*, 1109–1113.

Consortium, T.1000 G.P. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. PNAS *99*, 5261–5266.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic Gene Expression in a Single Cell. Science *297*, 1183–1186.

Fan, H.C., Wang, J., Potanina, A., and Quake, S.R. (2011). Whole-genome molecular haplotyping of single cells. Nature Biotechnology *29*, 51–57.

Futuyma, D. (2009). Evolution, Second Edition (Sinauer Associates, Inc.).

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Hanson, E.K., and Ballantyne, J. (2005). Whole genome amplification strategy for forensic genetic analysis using single or few cell equivalents of genomic DNA. Anal. Biochem. *346*, 246–257.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., et al. (2012). Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. Cell *148*, 873–885.

Kitzman, J.O., Snyder, M.W., Ventura, M., Lewis, A.P., Qiu, R., Simmons, L.E., Gammill, H.S., Rubens, C.E., Santillan, D.A., Murray, J.C., et al. (2012). Noninvasive Whole-Genome Sequencing of a Human Fetus. Sci Transl Med *4*, 137ra76–137ra76.

Krebs, J.E., Goldstein, E.S., and Kilpatrick, S.T. (2009). Lewin's Genes X (Jones & Bartlett Publishers).

Lao, K., Xu, N.L., and Straus, N.A. (2008). Whole genome amplification using single-primer PCR. Biotechnol J *3*, 378–382.

Lengauer, C., Kinzler, K.W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. Nature *396*, 643–649.

Li, G.-W., and Xie, X.S. (2011). Central dogma at the single-molecule level in living cells. Nature *475*, 308–315.

Lo, Y.M.D., Chan, K.C.A., Sun, H., Chen, E.Z., Jiang, P., Lun, F.M.F., Zheng, Y.W., Leung, T.Y., Lau, T.K., Cantor, C.R., et al. (2010). Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus. Sci Transl Med *2*, 61ra91–61ra91.

MacArthur, D. (2012). Methods: Face up to false positives. Nature *487*, 427–428.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nature Reviews Genetics *11*, 31–46.

Nagrath, S., Sequist, L.V., Maheswaran, S., Bell, D.W., Irimia, D., Ulkus, L., Smith, M.R., Kwak, E.L., Digumarthy, S., Muzikansky, A., et al. (2007). Isolation of rare circulating tumour cells in cancer patients by microchip technology. Nature *450*, 1235–1239.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Negrini, S., Gorgoulis, V.G., and Halazonetis, T.D. (2010). Genomic instability--an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. *11*, 220–228.

Project, the 1000 G. (2011). Variation in genome-wide mutation rates within and between human families. Nature Genetics *43*, 712–714.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Roach, J.C., Glusman, G., Smit, A.F.A., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science *328*, 636–639.

Rochette, P.J., Bastien, N., Lavoie, J., Guérin, S.L., and Drouin, R. (2005). SW480, a p53 double-mutant cell line retains proficiency for some p53 functions. J. Mol. Biol. *352*, 44–57.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science *239*, 487–491.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. (2004). Large-Scale Copy Number Polymorphism in the Human Genome. Science *305*, 525–528.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. Cell *144*, 27–40.

Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjöld, M., Ponder, B.A., and Tunnacliffe, A. (1992). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics *13*, 718–725.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. Cell *150*, 402–412.

Woyke, T., Sczyrba, A., Lee, J., Rinke, C., Tighe, D., Clingenpeel, S., Malmstrom, R., Stepanauskas, R., and Cheng, J.-F. (2011). Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. PLoS ONE *6*, e26161.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., et al. (2012). Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. Cell *148*, 886–895.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W., and Church, G.M. (2006). Sequencing genomes from single cells by polymerase cloning. Nature Biotechnology *24*, 680–686.

Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W., and Arnheim, N. (1992). Whole genome amplification from a single cell: implications for genetic analysis. Proc Natl Acad Sci U S A *89*, 5847–5851.

Chapter 7

Genome-wide Study of Meiotic Recombination in an Individual by Whole Genome Sequencing of Single Sperm Cells

7.1 Summary and Background

In Chapter 6, we introduced a whole genome amplification method that allows genome-wide analysis using individual cells. In this chapter, we explore a specific application of single cell genomics.

Meiotic recombination is essential to the proper segregation of homolog chromosomes, which results in the exchange of genetic information through crossover events and creates diversity

for evolution (Coop and Przeworski, 2006). Population analysis is widely used in studying human recombination, but yields results that are averaged among individuals and complicated by natural selection. Here we perform whole genome sequencing on 99 individual sperm cells from an Asian male using the newly developed MALBAC method described in Chapter 6. This allows us to construct a phased genome of the individual and determine the crossover positions in each sperm with high resolution, from which we build a personal map of recombination. We provide crucial evidence that the decrease in recombination rates near transcription start sites is intrinsic to the meiosis process, rather than due to selection. Furthermore, we find a significant propensity of autosomal aneuploidy with decreased crossover activity during spermatogenesis.

Meiosis plays a crucial role in sexual reproduction, in which homologue chromosomes are segregated to generate haploid gametes. This segregation requires establishment of physical connections and recombinations that result in crossovers between homologue chromosomes (Petronczki et al., 2003). Failure to form crossovers results in aneuploidy, which is the leading cause of miscarriage and birth defects (Epstein, 2007). Crossovers also create new combinations of alleles and contribute to genetic diversity and evolution (Jeffreys and May, 2004).

In human, Meiotic recombination has been mainly studied using population-based studies, such as by linkage disequilibrium (LD) of DNA markers (Ardlie et al., 2002; Myers et al.,

2005). The principle is shown in Figure 7.1.

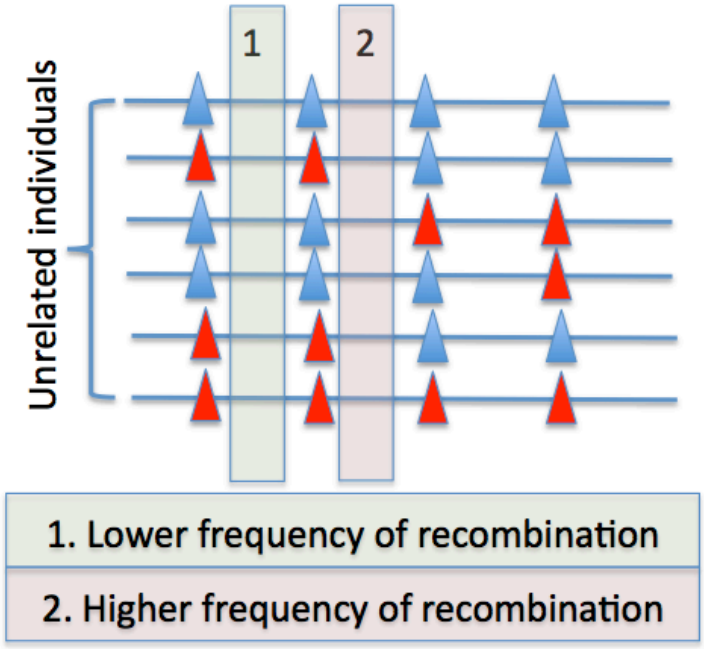


Figure 7.1 Linkage disequilibrium of DNA markers on the genome indicates different recombination frequencies differences across the human genome.

These population studies such as linkage disequilibrium (LD) and pedigree studies showed that the distribution of recombination is highly uneven across the human genome (Myers et al., 2005; Paigen and Petkov, 2010). Substantial recombination active regions are not conserved between humans and chimpanzees (Ptak et al., 2005; Winckler et al., 2005; Auton et al., 2012) and among different human populations (Kong et al., 2010; Hinch et al., 2011), suggesting these regions are quickly evolving and may even be individual-specific (Jeffreys and May, 2004).

This necessitates studying recombination genome-wide at the single human level, which could be in principle achieved by genotyping lots of offspring from the tested individual. However, unlike yeast and mouse, it is both unethical and infeasible to generate enough offspring for mapping recombinations. Another approach is to genotype individual gametes, but it requires a reliable whole genome amplification method for single cells. Such an effort will provide a deeper understanding of not only recombination and evolution, but also of the clinical relevance of recombination, such as in germline aneuploidy and infertility (Ferguson et al., 2007; Torres et al., 2008).

Despite the importance, studying meiotic recombination genome-wide at the individual level has been very challenging. Pedigree studies are often limited by the number of offspring (Kong et al., 2002) and sperm-typing experiments are often loci specific (Sarbajna et al., 2012). Whole genome amplification (WGA) of single sperm cells was proposed decades ago for mapping recombination (Zhang et al., 1992), and with the development of high throughput genotyping technologies (Lockhart and Winzeler, 2000; Metzker, 2010), whole-genome mapping of personal recombination events has become achievable and was recently demonstrated by performing WGA using Multiple displacement amplification (MDA) (Dean et al., 2002) on single sperm cells followed by genotyping using DNA microarray (Wang et al., 2012). However, due to the amplification unevenness and insufficient marker density, the resolution of crossover detection was limited to ~150kb on average. This study also relied on

prior knowledge of the somatic genome phase information, which is experimentally difficult to obtain and is currently available to only a few individuals (Levy et al., 2007; Peters et al., 2012; Wang et al., 2012).

In this chapter we demonstrate a more general approach to study recombination in single sperm cells of an individual, without assuming prior knowledge of the somatic phase information. The method is based on whole genome amplification and sequencing of single sperm cells, and is generally applicable to males from humans and other species.

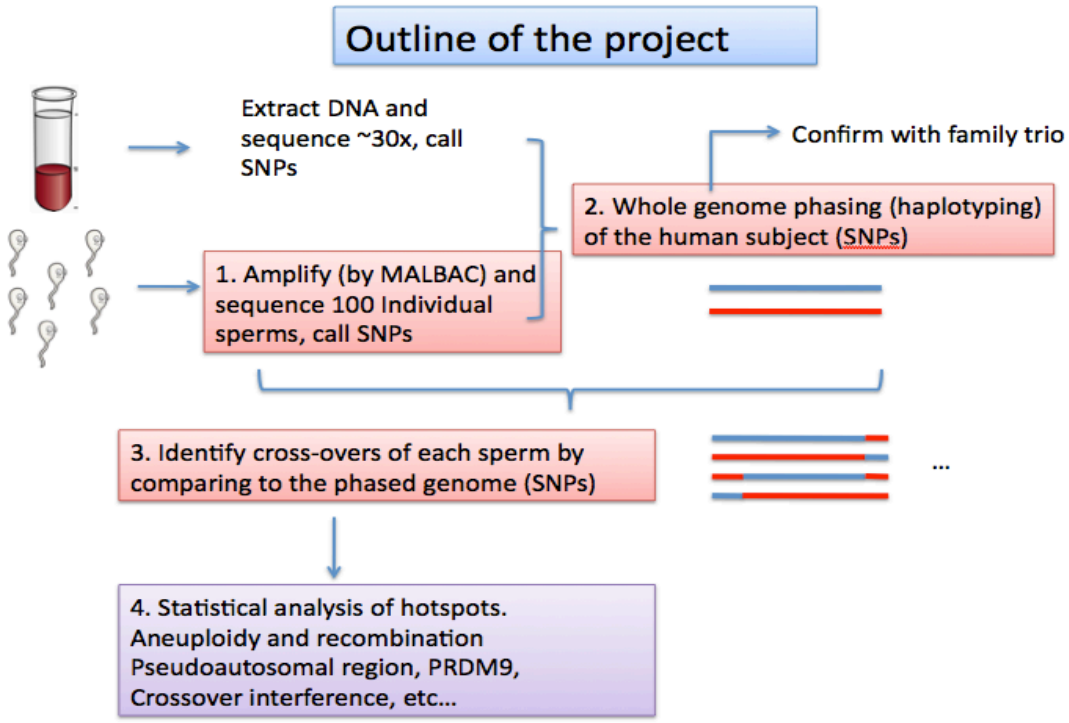


Figure 7.2 An outline of using whole genome sequencing of single sperm cells to study meiotic recombination of an individual.

An outline of the project is shown in Figure 7.2. We first amplify and sequence the whole genome of ~100 single sperm cells from an individual. We identify heterozygous SNPs (hetSNPs) from the bulk genome and their correspondent genotype from each single sperm, which allows us to build a phased-resolved human genome. We then identify the positions of crossovers and perform statistical analyses with the position information.

7.2 Whole Genome Amplification and Sequencing of Individual Sperm Cells

Experimentally, we first isolated single sperm cells from a normal Asian male donor at his late 40s. The donor has healthy offspring of both genders and normal clinical semen analysis result. The sperms were diluted to ~ 1/mm^2 using PBS+1%BSA on a petri-dish before mouth pipetting to isolate each single sperm into a reaction tube. The sperms were washed twice by PBS+1%BSA before lysed 3 hours in the Lysis Buffer as described previously in Chapter 6. We performed WGA on single cells using the recently developed method Multiple Annealing and Looping Based Amplification Cycles (MALBAC) described in Chapter 6. MALBAC provides significantly improved amplification evenness compared with the prevailing WGA methods, such as multiple displacement amplification (MDA) (Dean et al., 2002; Jiang et al., 2005; Lasken, 2007).

We generated ~2 micrograms of DNA for whole genome high throughput sequencing using an Illumina HiSeq 2000 sequencing platform. We sequenced 93 cells at ~1x depth and 6 cells at ~5x and achieved genome coverages of ~23% and ~43% respectively. 3 out of the 99 sperm samples were found to contain more than one haploid cell and were filtered out in downstream analysis (Figure 7.3). The average mapping rate of sequencing reads from single sperm is ~89%, which is close to that of a typical human resequencing project. This indicates the lack of spurious sequences from MALBAC amplification.

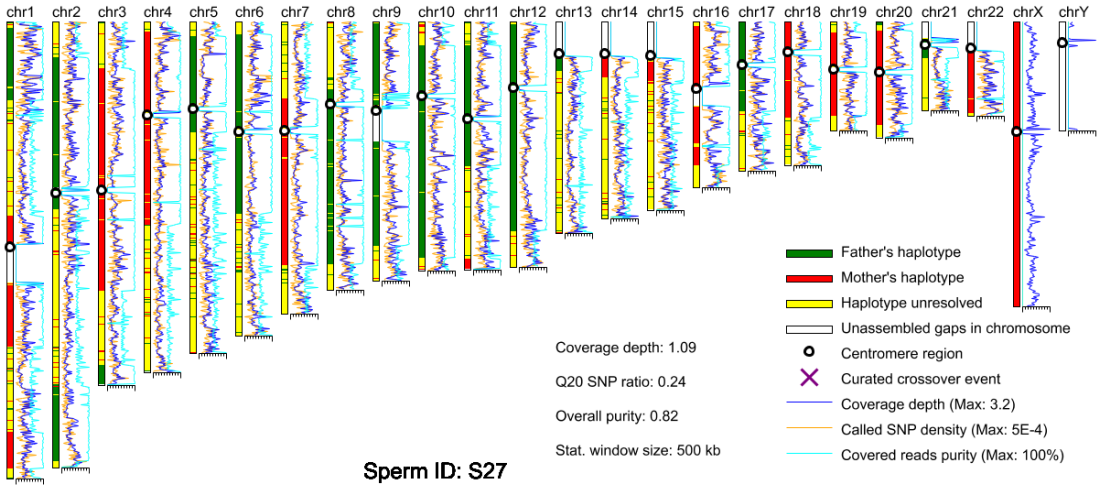


Figure 7.3: An example of the cells containing more than one haploid cell. There are multiple big genome fragments containing haplotypes that are contributed by both parents, indicating the analyzed sample contains more than a haploid genome. The 3 cells in this category are filter out to ensure accuracy in the following analyses.

We further sequenced the diploid genome of the donor at ~70x depth for obtaining a high quality personal genome. DNA molecules were extracted from freshly drawn blood using QIAGEN blood DNA extraction kit. About 10 microgram of DNA was extracted from the blood sample from the donor and his parents, which is then used for preparing sequencing libraries using standard Illumina protocol and were sequenced on an Illumina HiSeq 2000 sequencing platform.

We identified ~2.8 million single nucleotide polymorphisms (SNPs), ~1.4 million being heterozygous (hetSNPs). We then checked the genotype of each sperm on these sites. Among the hetSNP sites, ~500k (35%) and ~300k (20%) could be genotyped with Q20 threshold (error rate <1%) for the high coverage (5x) and low coverage (1x) sperm cells, respectively.

7.3 Whole Genome Haplotyping by Sequencing Individual Sperm

Phase information is crucial for the correct description and interpretation of the human genome (Bansal et al., 2011; Tewhey et al., 2011) and is essential for mapping crossovers. We phased the hetSNPs into chromosome-level haplotypes by comparing the SNP linkage information across all sperm cells (Figure 7.4).

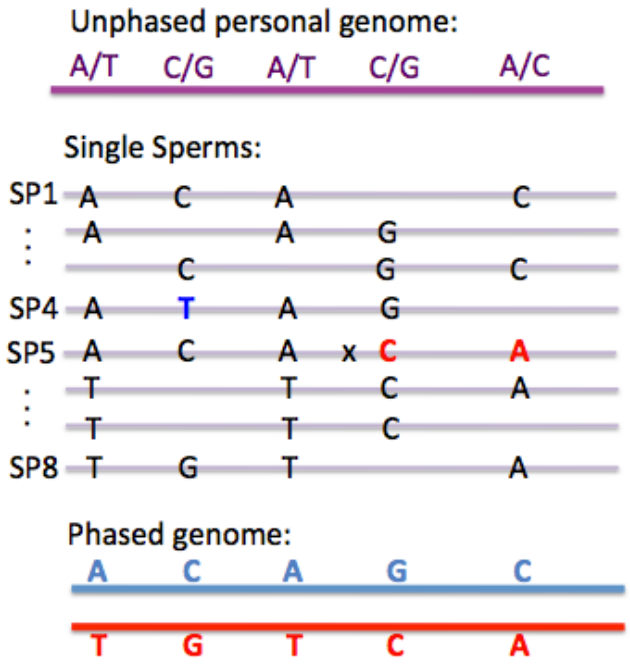


Figure 7.4: Principle of whole genome phasing of an individual using the SNP linkage information from individual sperm cells. We sequenced the diploid genome and identified five heterozygous SNPs with unknown linkage information shown in purple. Individual sperm cells were sequenced after MALBAC amplification, from which SNP linkage information in each sperm was used to infer the phase information in the diploid genome.

In brief, crossovers (such as the A-C link in SP5) and false SNP identification (such as the highlighted T in SP4) are low probability events, therefore most SNP linkage information identified in a sperm reflects the true SNP linkage in the somatic genome. These SNP linkages were calculated statistically by comparing across all sperm cells.

We developed a two-stage method to phase the diploid heterozygous SNPs (hetSNP) into chromosome-level haplotypes. First, we used a small subset (~10%) of the hetSNPs that are

covered by more than 40 sperm SNPs (quality >= Q20) to generate a framework of the haplotypes. The number of links for each of the 4 combinations was counted between all neighboring hetSNP sites. For example, if the genotype of one hetSNP is “CT”, and the genotype of its neighboring hetSNP is “AG”, then in one sperm, there would exists one of the “C-A” “C-G” “T-A” “T-G” links, in which two are true links and two others are false links. Assuming the rates of false SNP calling and recombination are low, the true links will appear much more frequently than the false links, which is the foundation of our method to infer the haplotype with sperms. We required at least 10 true links (the major type of links) for a neighboring hetSNP pair, and the number of false links (the minor type of links) must be less than 1/5 of the true links. The hetSNPs satisfying these criteria were phased into one of the two haplotypes. At the second stage, we attempted to fill in the other hetSNPs into the haplotype framework. For the sequencing reads covered a hetSNP site, we inspected five of the phased hetSNP sites (from the first stage) upstream and downstream. A hetSNP site can be phased if more than 80% of the inspected SNPs belong to the same haplotype. In addition, we require at least 2 sperms cover a hetSNP site, and there is no significant conflict in the result from different sperm cells (major# > 3x minor#).

By doing this, we phased ~1.1M (~82%) hetSNPs with high confidence into two sets of chromosome haplotypes. To verify the accuracy of phasing somatic genome using SNP linkage information in sperm, we lightly sequenced the genomes from the donor’s parents (~10x each) and inferred the SNPs from the parents with Q20 threshold. We took an

independent approach to phase the hetSNPs by comparing the genotypes of the donor and his parents. For example, if the genotype of a hetSNP site is C/T from the donor, then one of the parents must have contributed a "C", and the other parent must have contributed a "T". Except for the case of both parents being C/T, the hetSNP site can be phased into either paternal or maternal origin.

We obtained ~99.5% consistency of the two methods, indicating the high accuracy of our approach in phasing hetSNPs into chromosome-level haplotypes. We note that the percentage of phased hetSNPs can be further improved with higher sequencing depths from each sperm (currently only ~1x).

Several methods for haplotyping individual humans have been reported (Suk et al., 2011; Kitzman et al., 2012; Peters et al., 2012), However these methods often involve labor-intensive sample preparations such as cloning and have limited haplotype block size (<1Mb). Our method enables whole genome phasing into haplotypes of a complete chromosome, without requiring cell culture and sophisticated instrumentation or devices for separating metaphase chromosomes (Ma et al., 2010; Fan et al., 2011; Yang et al., 2011).

7.4 Crossover Distribution in Each Sperm

With the diploid genome phased into haplotypes of single chromosomes, we mapped the crossover positions of each sperm by identifying the SNP linkages crossing the border of the

two haplotypes. We used a hidden Markov model to accurately determine the positions for most crossovers and manually identified the crossovers for the remaining low confidence regions, and here we show an example of the identified crossovers from a sperm in Figure 7.5.

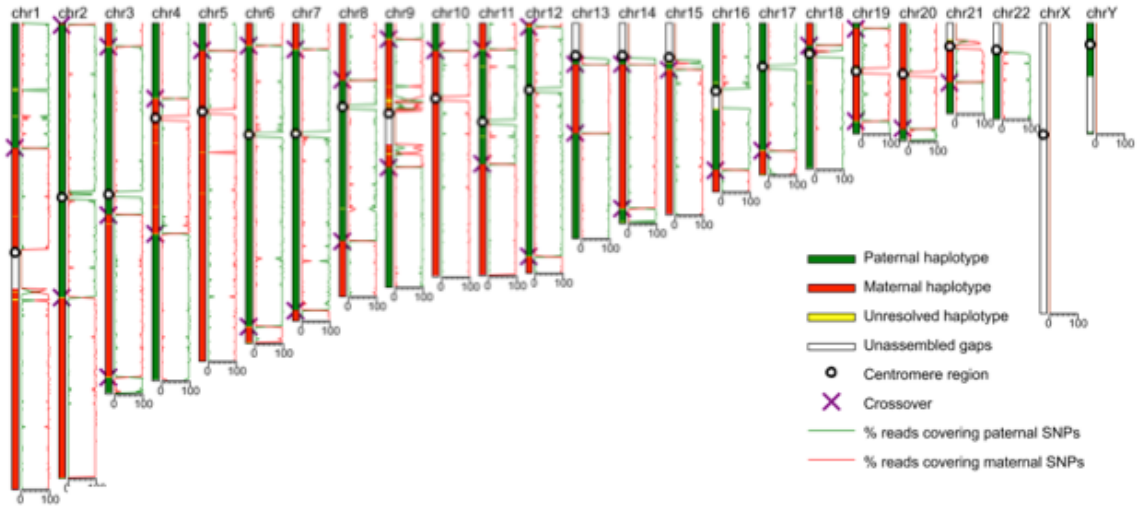


Figure 7.5: Identifying crossover positions in individual sperm cells. Parental haplotype contributions are determined by comparing the percentage of reads covering the paternal or maternal SNPs, and crossover positions are detected by identifying the crossing locations of the two parental haplotypes by a hidden Markov model.

We identified 2368 autosomal crossover events in the sperm cells that had a complete haploid genome. The average of ~26.0 crossovers per cell is consistent with the reported pedigree studies (Kong et al., 2002; Coop et al., 2008). With the improved amplification evenness of MALBAC, we achieved high resolution in detecting crossovers with only ~1x sequencing depth from each sperm, as is shown in Figure 7.6. About 93%, 80% and 45% of the crossovers can be confidently determined to intervals of 200 kb, 100 kb, and 30kb,

respectively, compared to 59%, 37% and 13% from the recently reported single sperm study (Wang et al., 2012). This resolution is also significantly better than some pedigree studies (Kong et al., 2002; Myers et al., 2005). Of the crossovers resolvable within a 10kb interval, ~40% of the crossovers are found to overlap with the male-specific recombination hotspots inferred from the deCODE project (Kong et al., 2010). About 45% of the crossovers are close to the PRDM9 binding motif CCnCCnTnnCCnC (Berg et al., 2010; Parvanov et al., 2010), which is consistent with the previous studies in the population (Coop et al., 2008; Baudat et al., 2010; Kong et al., 2010).

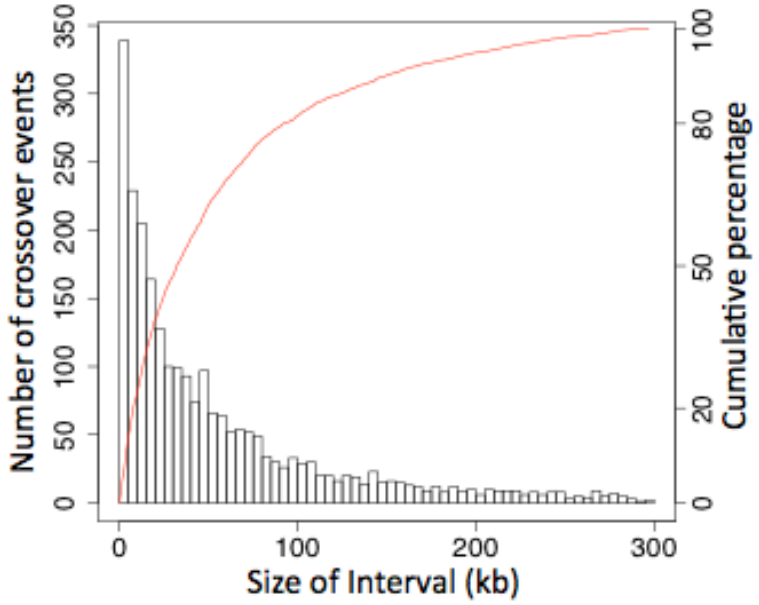


Figure 7.6: Resolution of crossover determination. Here shows the number and the cumulative fraction of the crossovers that can be resolved into a certain size of interval. ~60% of the crossovers can be determined within intervals of 50kb.

The high resolution of crossover detection in sperm cells allows detailed inspection of local recombination features with negligible selection effect. It is known that recombination rates correlate positively with gene density both in yeast and human (Petes, 2001; Coop et al., 2008) . However, at a finer scale, recombination rates are observed to be actually lower close to genes and higher tens or hundreds of kilobase away from the transcription start sites (TSS) in the population (Myers et al., 2005; Coop et al., 2008; Kong et al., 2010). However, it is not clear whether this distinct feature is general to all human beings and really reflects the recombination variation during meiosis or is mainly an effect of selection. With the ability to precisely identify crossovers in single sperm cells, we derived the recombination rate relative to the TSS of the individual directly without selection effect. We included only the crossovers resolvable within 30kb for the analysis.

As shown in Figure 7.7, we compared the results from our sperm data and from a previous population study (Coop et al., 2008). The recalculated the male-specific recombination distribution in the previous study (Coop et al., 2008).

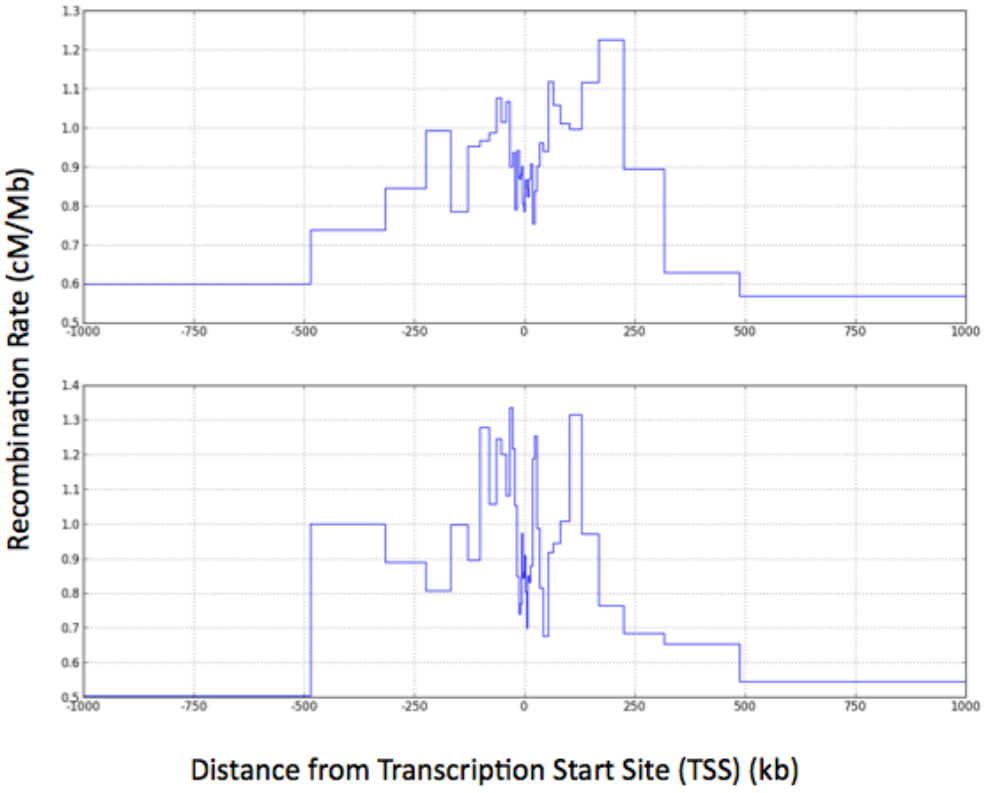


Figure 7.7: Distribution of recombination rate relative to transcription start sites (TSS). Recombination rates are recalculated from (Coop et al., 2008) on male individuals only (top panel). We compare this to the recombination rates derived from the sperm data (bottom panel).

We observed lower recombination rate close to the TSS and higher rate tens of kilobase away from the TSS, which is consistent with the previous population studies (Myers et al., 2005; Coop et al., 2008; Kong et al., 2010), indicating the lower recombination rate close to TSS is primarily due to the variation of recombination probability during meiosis rather than selection on variations of offspring viability.

7.5 Genome-wide Distribution of Recombination

Recombination events are known to have a non-uniform distribution across the genome by previous population studies (Petes, 2001; Paigen and Petkov, 2010). By binning the crossover incidence into units of three megabases in autosomes, we constructed a personal genetic map of recombination, and we compared it to a population-based sex-averaged map (HapMap) (Myers et al., 2005) and a pedigree-based male-specific map (deCODE) (Kong et al., 2010) (Figure 7.8).

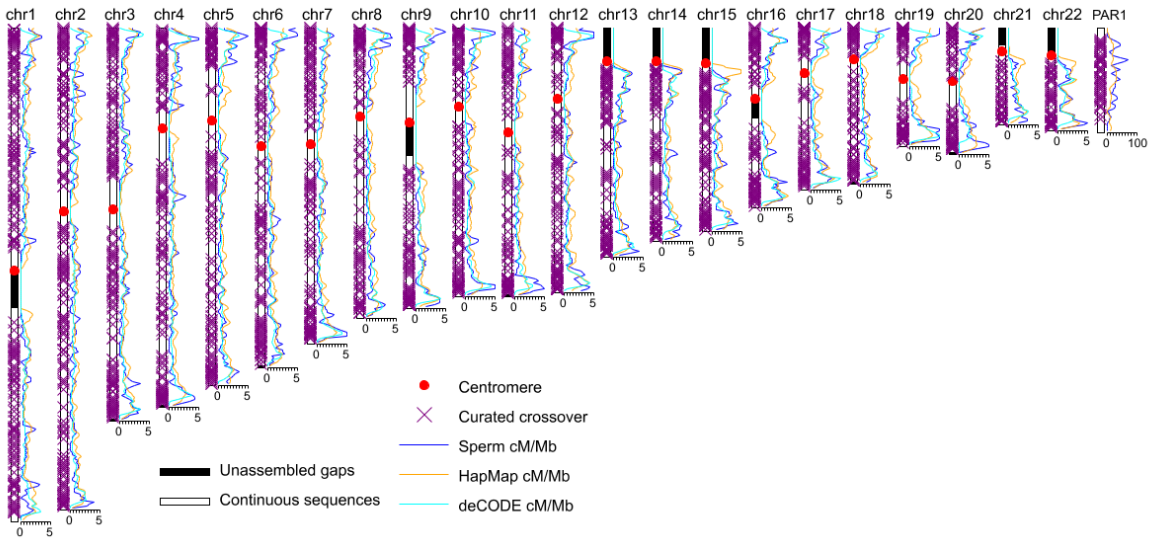


Figure 7.8: Genome-wide distribution of recombination (A) Comparison of the sperm recombination rates to the HapMap and deCODE (male-specific) genetic maps across the human genome. We used a 3Mb statistical window size and a 1Mb moving step.

We obtained correlation coefficients of 0.71 and 0.77 of the sperm derived recombination

rates with HapMap and DeCODE respectively. We also plot out the cumulative distribution of recombinations along a chromosome (Figure 7.9). The map derived from HapMap shows significant deviation from the other two because female has significantly higher recombination rate than male in general, and the rate of recombination inferred from HapMap is sex-averaged.

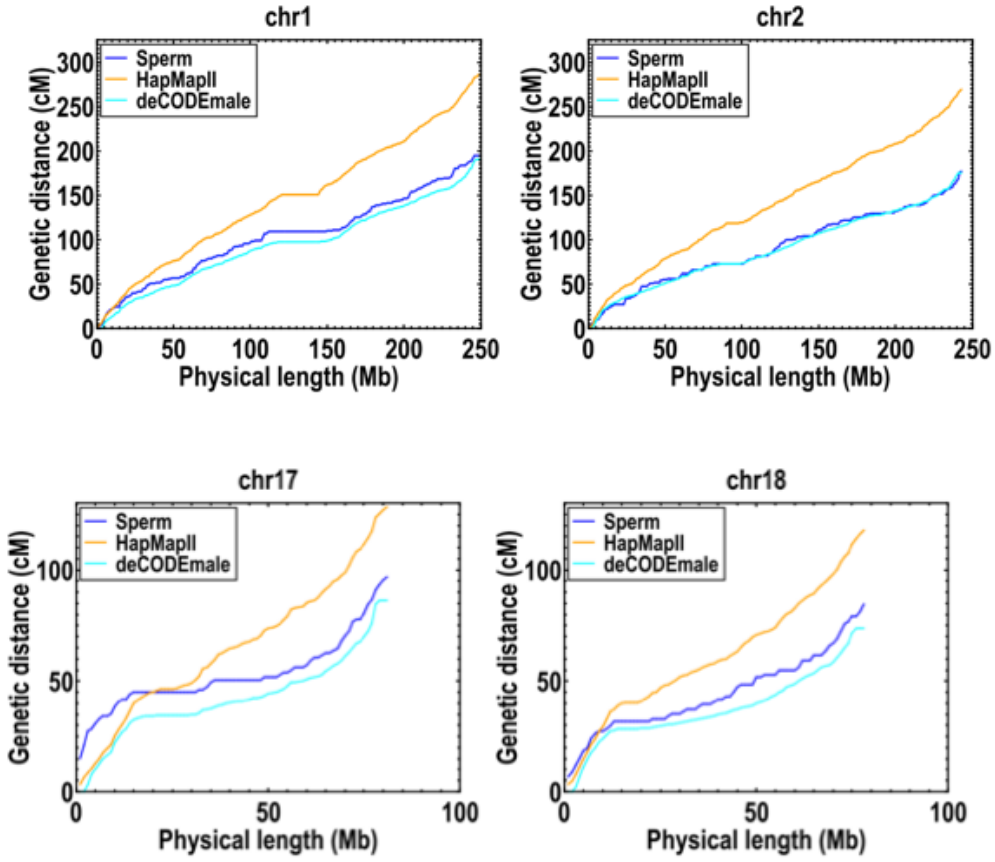


Figure 7.9: A personal genetic map of recombination. Relations of physical and genetic length of selected chromosomes

In some of the 3Mb statistical bins, we observed a significant difference between HapMap and the donor (Table 7.1). They can be explained by sex-specific recombination variations (Kong et al., 2010).

Chr ID	Start	End	Crossover events	Sperm recombination rate	HapMap recombination rate	P-value
Chr12	3,000,001	6,000,000	26	28.57%	8.57%	6.3×10^{-6}
Chr19	54,000,001	57,000,000	23	25.27%	11.47%	6.9×10^{-2}

Table 7.1: 3Mb bins that showed significant difference in recombination rate between population (HapMap) and the donor (inferred by crossovers in sperm). These differences can be explained as sex-specific recombination active regions. The P-values were corrected for multiple testing (~1000 tests).

A previous study reported crossover active regions that are specific to an individual exist at a megabase scale (Wang et al., 2012). Such finding, if true, would imply extraordinary rapid evolution of human recombination across the genome, even at a large megabase scale. Indeed, we also found 9 bins showing significant differences between the donor and deCODE (Table 7.2).

Chr ID	Start	End	Crossover events	Sperm recombination rate	deCODE recombination rate	P-value
Chr3	195,000,001	198,000,000	5	5.49%	0.33%	$6.8x10^{-4}$
Chr9	138,000,001	141,000,000	11	12.09%	3.05%	$2.1x10^{-2}$
Chr11	132,000,001	135,000,000	9	9.89%	0.98%	$2.6x10^{-5}$
Chr14	21,000,001	24,000,000	9	9.89%	2.12%	$2.5x10^{-2}$
Chr14	63,000,001	66,000,000	7	7.69%	1.26%	$2.2x10^{-2}$
Chr20	57,000,001	60,000,000	12	13.19%	0.05%	$1.6x10^{-12}$
Chr20	60,000,001	63,000,000	9	9.89%	0%	0
Chr21	45,000,001	48,000,000	9	9.89%	0.61%	$2.8x10^{-7}$
Chr21	48,000,001	51,000,000	12	13.19%	1.90%	$2.1x10^{-5}$

Table 7.2: 3Mb bins that showed significant difference in recombination rate between population (deCODE male specific) and the donor (inferred by crossovers in sperm). The first and last 3MB was removed from the comparison for poor marker density in deCODE. The P-values were corrected for multiple testing (~1000 tests). We note that most of these regions are at the ends or close to the centromere region in which the accuracy for inferring crossovers is low in deCODE. Therefore, we tend not to interpret them as being personal recombination active regions.

However, we note that most of these regions are very close to the centromere or the ends of the chromosomes, where the estimation of the recombination rates was considered not reliable and excluded in deCODE (Kong et al., 2010). Therefore, we suspect these differences mainly reflect the incompleteness of the deCODE database. Our results suggest the distribution of

recombination in the individual generally agrees with the population average at a megabase scale, which indicates a general consistency of large-scale recombination distribution in human evolution.

Although we did not see large-scale variations between the individual and the population. We cannot exclude the possibility that small-scale variations (such as in several kilobases) exist between different individuals. To see variations at smaller scale requires sequencing more sperm from the individual. With the rapid development of sequencing technologies, such effort is becoming feasible financially in the near future, and more sperm can be analyzed from different individuals to look into fine-scale recombination variations in a population, or in different pathological states. We estimated the number of sperm required for seeing these differences with statistical significance in Figure 7.10.

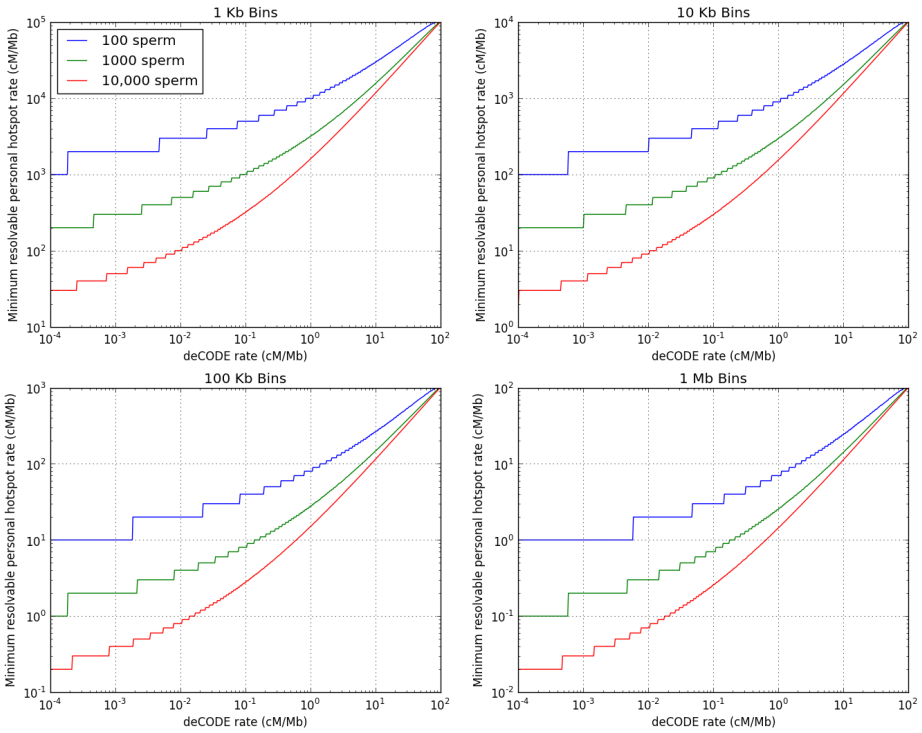


Figure 7.10: Estimations of the number of sperm required for detecting the differences on recombination rates, between an individual and the population average if such differences exist. For example, if the recombination rate of deCODE is 0.1cM/Mb in a 1Mb window, the recombination rate of the individual will need to be ~0.9cM/Mb in order to be detected from 1000 sperms with statistical significance (P=0.05, corrected for multiple testing).

7.6 Pseudoautosomal Region and Crossover Interference

Obtaining the genome sequence of each sperm also allows us to examine the crossover events in the pseudoautosomal region (PAR) of the sex chromosomes (Figure 7.11). Human PAR contains ~3 megabases of homologous sequences that are located in both ends of the sex chromosomes PAR is essential in establishing crossover of the sex chromosomes in males,

due to the requirement of sequence similarity for homologue recombination (Flaquer et al., 2008; Otto et al., 2011).

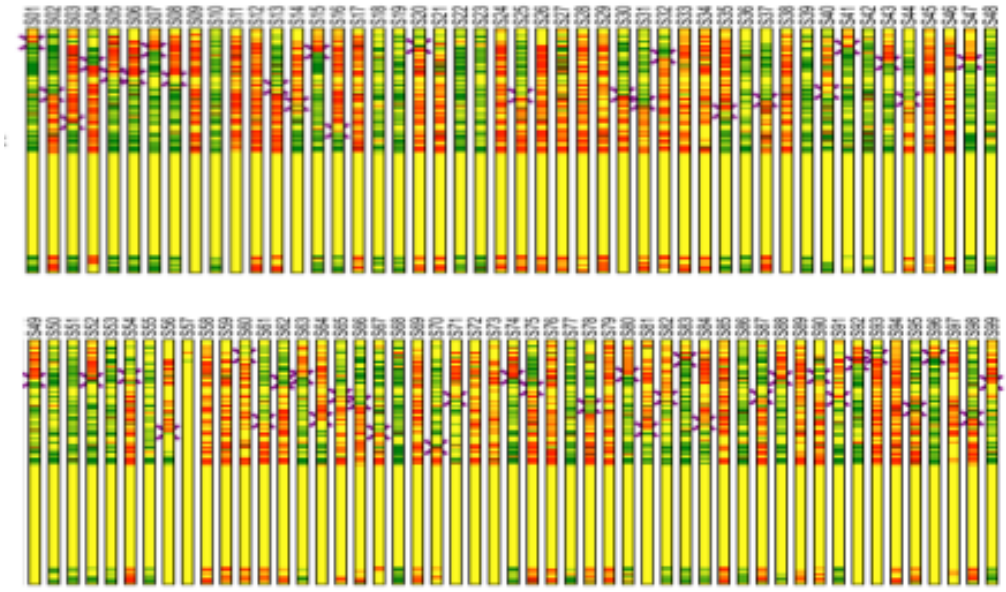


Figure 7.11: Crossover identified in the Pseudoautosomal Region (PAR) of each sperm

PAR contains limited microarray markers and a high percentage of repetitive sequences and was therefore excluded in previous pedigree and single sperm studies. By identifying the parental contribution of the hetSNPs in the uniquely mappable regions, we determined the crossover positions in PAR for each sperm. We detected on average ~0.6 crossovers per sperm. Interestingly, we did not see any two crossovers coexist on the same sperm, indicating a crossover tends to avoid the occurrence of another crossover in proximity, which is

consistent with the phenomenon known as crossover interference (Broman and Weber, 2000; Kleckner et al., 2004). Such an effect is also reflected in our observation on autosomes that crossovers tend to be separated by longer distance than would have been expected by random chance shown in Figure 7.12.

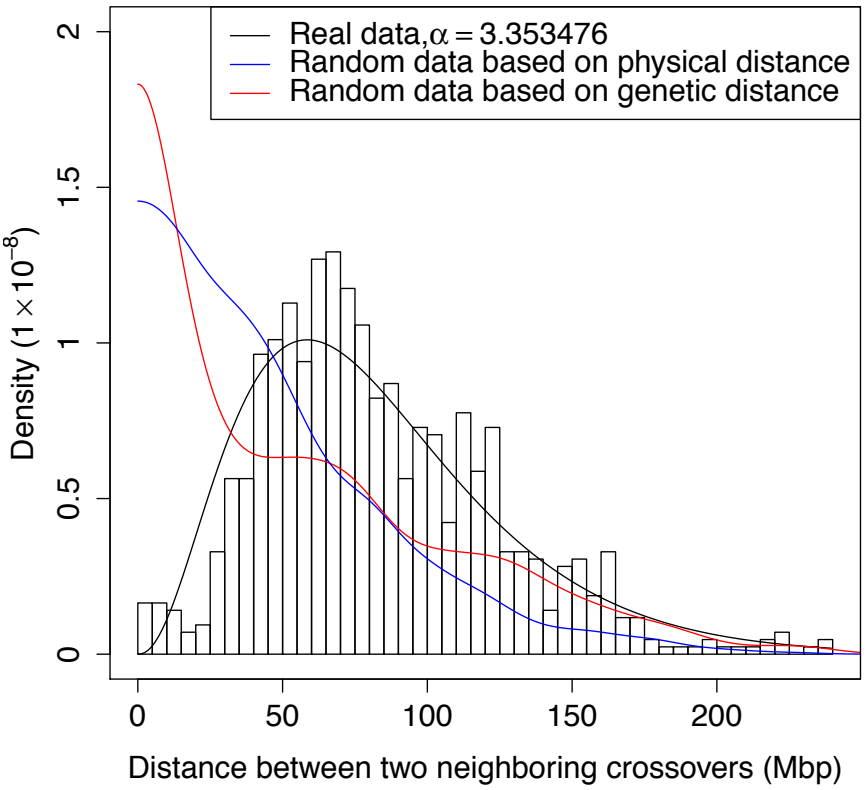


Figure 7.12: Distribution of the distance of coexisted crossover events on a chromosome. The histogram is fitted with gamma distribution and we obtain coefficient α~3.35. The significant deviation from random distribution indicates a crossover event tend to avoid other crossover events from happening in its proximity. In comparison, we generated random crossovers based on physical and genetic distances.

In humans, crossovers (CO) exhibit positive interference along chromosomes, resulting in more evenly spaced distribution than would be expected from random distribution (Broman and Weber, 2000; Kotwaliwale, 2012). Whole genome mapping of the CO sites from multiple sperm cells allows direct genome-wide study of the CO interference effect (Figure 7.12). Out of the 2420 CO events we identified, 819 pairs coexist on the same chromosome. We plotted the number of events with the physical distances between two COs and fitted with a gamma distribution. We observed a substantial deviation from the random distribution (α=3.35, random distribution α=1), indicating a CO tends to avoid the occurrence of another CO in proximity. We further compare the gamma distribution coefficient α of different chromosomes (Figure 7.13).

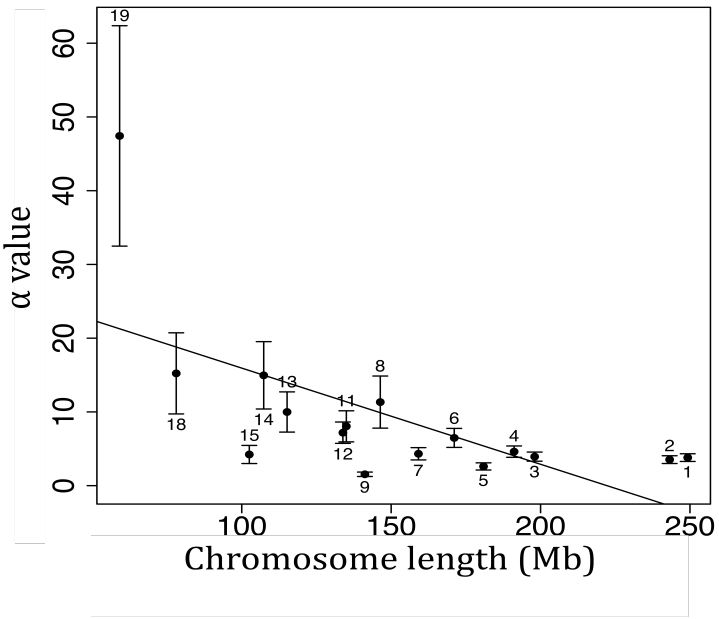


Figure 7.13: Comparison of the crossover interference effects on different chromosomes, the α value shows negative correlation with the size of the chromosome.

α shows the deviation from random distribution and can be used as a quantitative measurement of the strength of CO interference. We notice an anti-correlation of α with chromosome size, indicating a stronger CO interference effect with the shorter chromosomes, which is consistent with a previous report using fluorescence labeling (Lian et al., 2008). We further inspected the crossover events coexisted across the centromere, and we confirm that the interference of crossovers exist across the centromere.

Crossover interference:

crossing the centromere: not crossing the centromere:

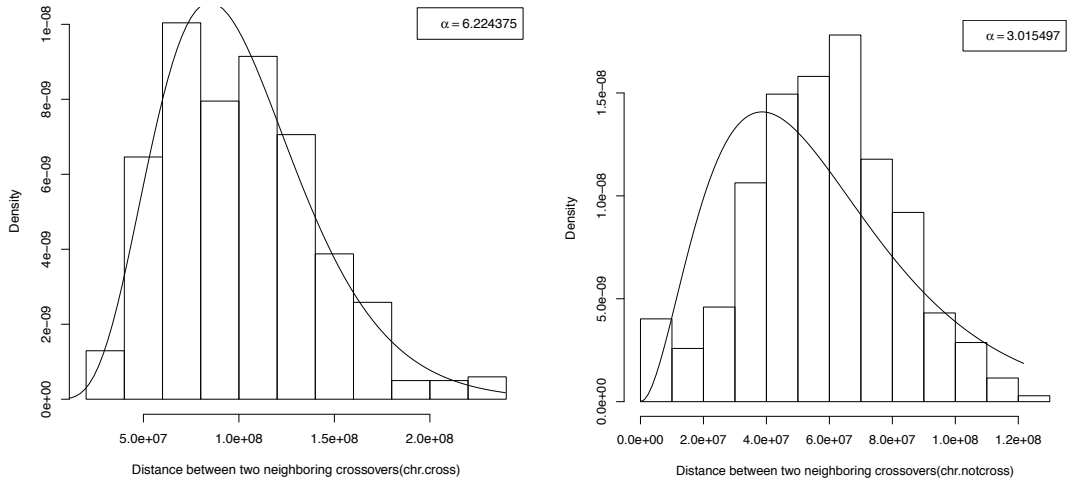


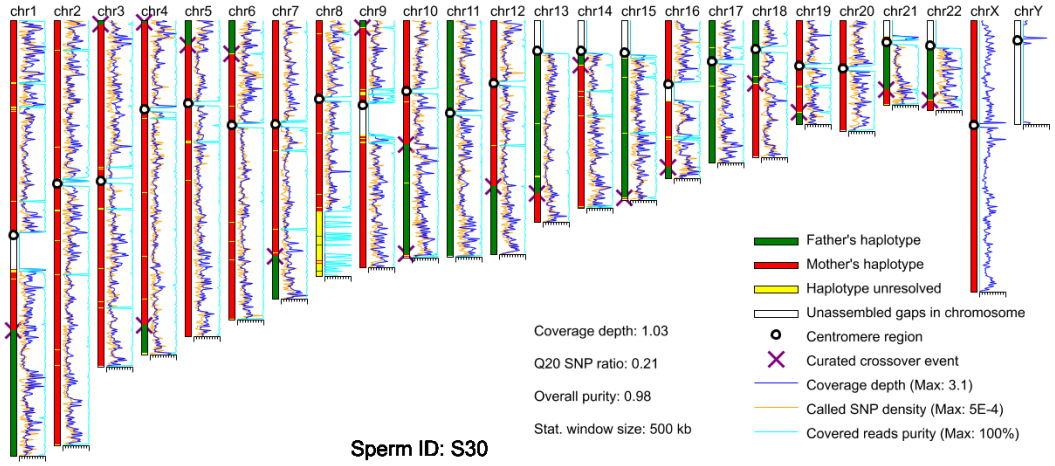
Figure 7.14: Crossover interference through centromere. Here we should the gamma distribution fit for Crossover pairs that are separated by centromere compared with those that are not separated by centromere. Significant deviation from random distribution were observed in both cases, indicating Crossover interference exists through centromere.

A previous study reported that substantial double crossovers occurring close together (e.g. 1-5 Mb) (Fledel-Alon et al., 2009). Although we have much higher resolution, we did not see either on autosomes or sex chromosomes, suggesting such phenomenon is likely not general and may only exist in certain populations.

7.7 Chromosome Segregation Error and Crossover

Failure of forming crossovers during meiosis gives rise to chromosome segregation errors that result in aneuploidy. Autosomal aneuploidy is often lethal to embryos, with the exception of a few chromosomes that result in severe health consequences early in development (i.e. Trisomy 21, Down Syndrome) (Gardiner et al., 2000). Reduced recombination activity is often found to associate with male infertility and sperm aneuploidy in pathological conditions (Ferguson et al., 2007). However, it is not clear whether in the sperm cells from males with normal fertility, chromosome aneuploidy is associated with reduced recombination activity. Whole genome sequencing allows simultaneously detecting chromosome aneuploidy with recombination positions. By comparing the coverage depth and SNPs along the genome of the sperm cells, we detected four cells either missing or having additional autosomes (Figure 7.15).

S30: Missing a large terminal fragment in chr8



S39: Missing chr 19

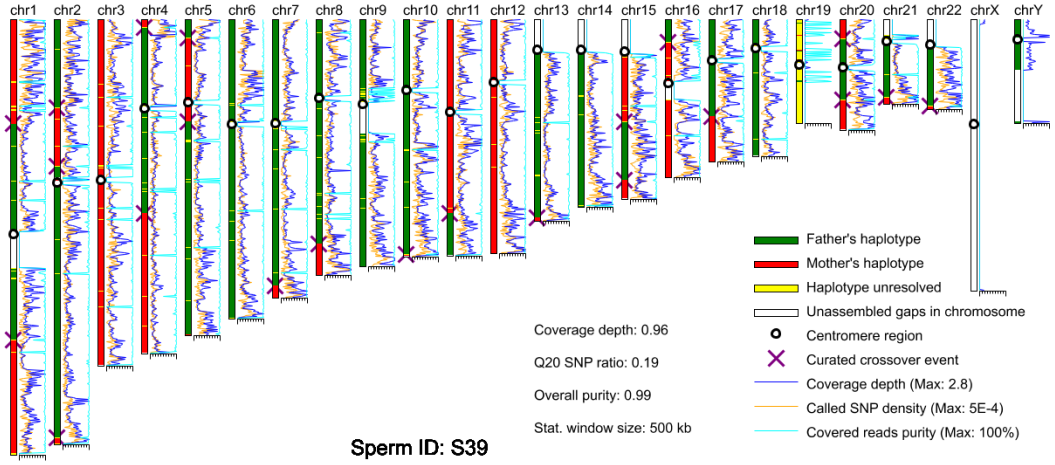
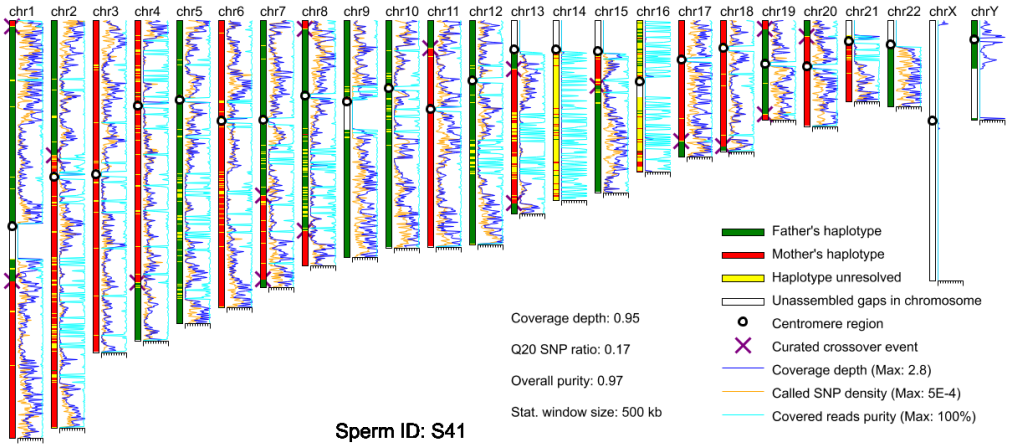


Figure 7.15 Sperm cells that show aneuploidy in autosomes. The four samples show very distinct pattern where only one or two chromosomes are abnormal while all other chromosomes have high purity of parental haplotype contributions.

S41: Missing chr14 and chr16



S65: Having an additional chr6:

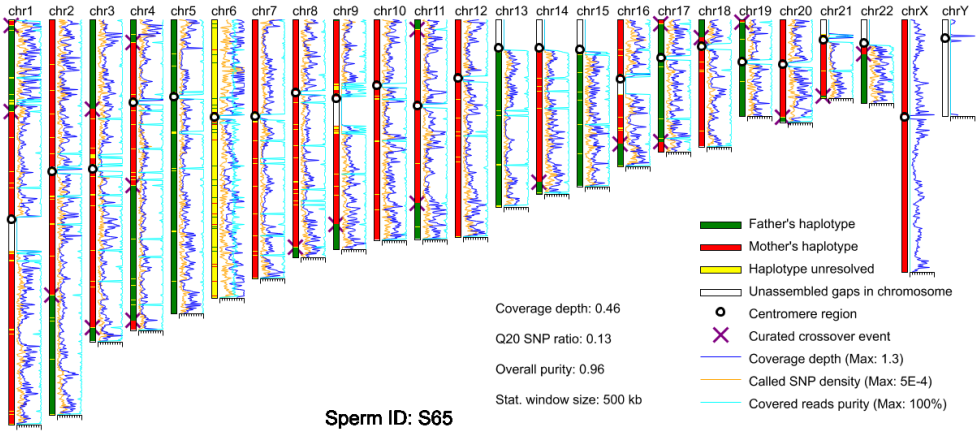


Figure 7.15 (continue): Sperm cells that show aneuploidy in autosomes.

The rate of chromosome mis-segregation in the sperm cells is consistent with the reported imaging studies on selected loci of human spermatocytes (Spriggs et al., 1995; Downie et al.,

1997). We then compared the crossover number of the aneuploid sperm cells to the normal group, the result is shown in Figure 7.16.

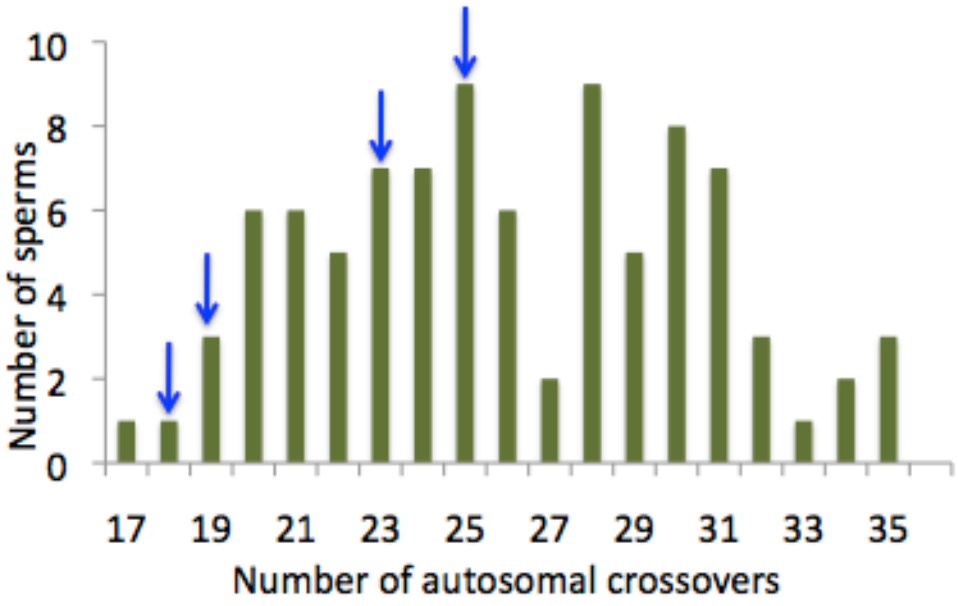


Figure 7.16. Detecting aneuploidy and crossover in the same sperm. Distribution of the autosomal crossover number of the normal cells. Blue arrows indicate the number of crossover in sperm cells with autosomal aneuploidy.

Interestingly, sperm cells with aneuploid autosomes exhibit significantly fewer crossovers than normal cells on average ($p=0.01$). Our result suggests that autosomal segregation error is not generated randomly during spermatogenesis. Instead, the error rate is higher in the spermatocytes with relatively repressed crossover activity. However, such a trend does not seem to be significant for sex chromosome aneuploidy, as we observed a sperm with 30 autosomal crossovers but no sex chromosome. Indeed, the crossover probability in the PAR

region of the sex chromosomes has no noticeable correlation with that of the autosomes (Table 7.3), suggesting a different mechanism of crossover generation for autosomes and sex chromosomes, which is consistent with an earlier study in mice (Kauppi et al., 2011).

Sperm ID	Autosome	PAR	Sperm ID	Autosome	PAR	Sperm ID	Autosome	PAR	Sperm ID	Autosome	PAR
S01	34	1	S24	21	0	S51	22	0	S77	25	0
S02	24	1	S25	21	1	S52	31	1	S78	23	1
S03	20	1	S26	26	0	S53	31	0	S79	35	0
S04	24	1	S28	23	0	S54	28	1	S80	24	1
S05	25	1	S29	24	0	S55	22	0	S81	19	1
S06	25	1	S31	34	1	S56	26	1	S82	23	1
S07	21	1	S32	30	1	S58	28	0	S83	21	1
S08	22	1	S33	20	0	S59	27	0	S84	30	1
S09	29	0	S34	28	0	S60	17	0	S85	32	0
S10	31	0	S35	23	1	S61	23	1	S86	25	0
S11	25	0	S36	25	0	S62	28	1	S88	33	1
S12	32	0	S37	21	1	S63	25	1	S89	28	0
S13	28	1	S38	29	0	S64	30	1	S90	29	1
S14	30	1	S40	28	1	S66	30	1	S91	23	1
S15	20	1	S42	35	0	S67	23	1	S92	31	1
S16	24	1	S43	28	1	S69	19	0	S93	31	1
S17	22	0	S44	24	1	S70	27	1	S94	29	0
S18	21	0	S45	26	0	S71	31	1	S95	20	1
S19	26	0	S46	29	0	S72	25	0	S96	22	1
S20	32	1	S47	26	1	S73	20	0	S97	28	0
S21	31	1	S48	19	0	S74	35	1	S98	30	1
S22	25	0	S49	26	1	S75	30	1	S99	18	1
S23	24	0	S50	30	0	S76	20	0			

Table 7.3: Crossover numbers called in each sperm. The average autosomal crossover number for sperm cells with crossover in PAR is 26.2, compared to 26.0 for all sperm cells. The difference is not statistically significant, indicating there is no noticeable correlation of crossover frequency between autosomes and sex chromosomes.

In summary, our approach of whole genome amplifying and sequencing single sperm cells from an individual enables direct examination of simultaneous chromosome segregation error together with crossovers in single sperm cells, in a whole genome, non-invasive and label-free manner.

References:

Ardlie, K.G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics *3*, 299–309.

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Ségurel, L., Street, T., Leffler, E.M., Bowden, R., Aneas, I., Broxholme, J., et al. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. Science *336*, 193–198.

Bansal, V., Tewhey, R., Topol, E.J., and Schork, N.J. (2011). The next phase in human genetics. Nature Biotechnology *29*, 38–39.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and Massy, B. de (2010). PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. Science *327*, 836–840.

Berg, I.L., Neumann, R., Lam, K.-W.G., Sarbajna, S., Odenthal-Hesse, L., May, C.A., and Jeffreys, A.J. (2010). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat. Genet. *42*, 859–863.

Broman, K.W., and Weber, J.L. (2000). Characterization of human crossover interference. Am J Hum Genet *66*, 1911–1926.

Coop, G., and Przeworski, M. (2006). An evolutionary view of human recombination. Nature Reviews Genetics *8*, 23–34.

Coop, G., Wen, X., Ober, C., Pritchard, J.K., and Przeworski, M. (2008). High-Resolution Mapping of Crossovers Reveals Extensive Variation in Fine-Scale Recombination Patterns Among Humans. Science *319*, 1395–1398.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. PNAS *99*, 5261–5266.

Downie, S.E., Flaherty, S.P., Swann, N.J., and Matthews, C.D. (1997). Estimation of Aneuploidy for Chromosomes 3, 7, 16, X and Y in Spermatozoa from 10 Normospermic Men Using Fluorescence in-Situ Hybridization. Mol. Hum. Reprod. *3*, 815–819.

Epstein, C.J. (2007). The consequences of chromosome imbalance: principles, mechanisms, and models (Cambridge Univ Pr).

Fan, H.C., Wang, J., Potanina, A., and Quake, S.R. (2011). Whole-genome molecular haplotyping of single cells. Nature Biotechnology *29*, 51–57.

Ferguson, K.A., Wong, E.C., Chow, V., Nigro, M., and Ma, S. (2007). Abnormal meiotic recombination in infertile men and its association with sperm aneuploidy. Hum. Mol. Genet. *16*, 2870–2879.

Flaquer, A., Rappold, G.A., Wienker, T.F., and Fischer, C. (2008). The human pseudoautosomal regions: a review for genetic epidemiologists. Eur. J. Hum. Genet. *16*, 771–779.

Fledel-Alon, A., Wilson, D.J., Broman, K., Wen, X., Ober, C., Coop, G., and Przeworski, M. (2009). Broad-Scale Recombination Patterns Underlying Proper Disjunction in Humans. PLoS Genet *5*, e1000658.

Gardiner, K., Davisson, M., and others (2000). The sequence of human chromosome 21 and implications for research into Down syndrome. Genome Biol *1*, 0002–1.

Hinch, A.G., Tandon, A., Patterson, N., Song, Y., Rohland, N., Palmer, C.D., Chen, G.K., Wang, K., Buxbaum, S.G., Akylbekova, E.L., et al. (2011). The landscape of recombination in African Americans. Nature *476*, 170–175.

Jeffreys, A.J., and May, C.A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat. Genet. *36*, 151–156.

Jiang, Z., Zhang, X., Deka, R., and Jin, L. (2005). Genome amplification of single sperm using multiple displacement amplification. Nucleic Acids Res *33*, e91.

Kauppi, L., Barchi, M., Baudat, F., Romanienko, P.J., Keeney, S., and Jasin, M. (2011). Distinct Properties of the XY Pseudoautosomal Region Crucial for Male Meiosis. Science *331*, 916–920.

Kitzman, J.O., Snyder, M.W., Ventura, M., Lewis, A.P., Qiu, R., Simmons, L.E., Gammill, H.S., Rubens, C.E., Santillan, D.A., Murray, J.C., et al. (2012). Noninvasive Whole-Genome Sequencing of a Human Fetus. Sci Transl Med *4*, 137ra76–137ra76.

Kleckner, N., Zickler, D., Jones, G.H., Dekker, J., Padmore, R., Henle, J., and Hutchinson, J. (2004). A mechanical basis for chromosome function. PNAS *101*, 12592–12597.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat. Genet. *31*, 241–247.

Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.

Kotwaliwale, C.V. (2012). Robustness in crossover regulation during meiosis. Nature Cell Biology *14*, 335–337.

Lasken, R.S. (2007). Single-cell genomic sequencing using Multiple Displacement Amplification. Current Opinion in Microbiology *10*, 510–516.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.

Lian, J., Yin, Y., Oliver-Bonet, M., Liehr, T., Ko, E., Turek, P., Sun, F., and Martin, R.H. (2008). Variation in crossover interference levels on individual chromosomes from human males. Hum. Mol. Genet. *17*, 2583–2594.

Lockhart, D.J., and Winzeler, E.A. (2000). Genomics, gene expression and DNA arrays. Nature *405*, 827–836.

Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K., and Song, Q. (2010). Direct determination of molecular haplotypes by chromosome microdissection. Nature Methods *7*, 299–301.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nature Reviews Genetics *11*, 31–46.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A Fine-Scale Map of Recombination Rates and Hotspots Across the Human Genome. Science *310*, 321–324.

Otto, S.P., Pannell, J.R., Peichel, C.L., Ashman, T.-L., Charlesworth, D., Chippindale, A.K., Delph, L.F., Guerrero, R.F., Scarpino, S.V., and McAllister, B.F. (2011). About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. Trends in Genetics *27*, 358–367.

Paigen, K., and Petkov, P. (2010). Mammalian recombination hot spots: properties, control and evolution. Nature Reviews Genetics *11*, 221–233.

Parvanov, E.D., Petkov, P.M., and Paigen, K. (2010). Prdm9 Controls Activation of Mammalian Recombination Hotspots. Science *327*, 835–835.

Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., et al. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature *487*, 190–195.

Petes, T.D. (2001). Meiotic recombination hot spots and cold spots. Nature Reviews Genetics *2*, 360–369.

Petronczki, M., Siomos, M.F., and Nasmyth, K. (2003). Un Ménage à Quatre: The Molecular Biology of Chromosome Segregation in Meiosis. Cell *112*, 423–440.

Ptak, S.E., Hinds, D.A., Koehler, K., Nickel, B., Patil, N., Ballinger, D.G., Przeworski, M., Frazer, K.A., and Pääbo, S. (2005). Fine-scale recombination patterns differ between chimpanzees and humans. Nature Genetics *37*, 429–434.

Sarbajna, S., Denniff, M., Jeffreys, A.J., Neumann, R., Artigas, M.S., Veselis, A., and May, C.A. (2012). A major recombination hotspot in the XqYq pseudoautosomal region gives new insight into processing of human gene conversion events. Hum. Mol. Genet. *21*, 2029–2038.

Spriggs, E.L., Rademaker, A.W., and Martin, R.H. (1995). Aneuploidy in human sperm: results of two-and three-color fluorescence in situ hybridization using centromeric probes for chromosomes 1, 12, 15, 18, X, and Y. Cytogenet. Cell Genet. *71*, 47–53.

Suk, E.-K., McEwen, G.K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D.T., McLaughlin, S., Peckham, H., et al. (2011). A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res *21*, 1672–1685.

Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J., and Schork, N.J. (2011). The importance of phase information for human genomics. Nature Reviews Genetics *12*, 215–223.

Torres, E.M., Williams, B.R., and Amon, A. (2008). Aneuploidy: cells losing their balance. Genetics *179*, 737–746.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. Cell *150*, 402–412.

Winckler, W., Myers, S.R., Richter, D.J., Onofrio, R.C., McDonald, G.J., Bontrop, R.E., McVean, G.A.T., Gabriel, S.B., Reich, D., Donnelly, P., et al. (2005). Comparison of Fine-Scale Recombination Rates in Humans and Chimpanzees. Science *308*, 107–111.

Yang, H., Chen, X., and Wong, W.H. (2011). Completely phased genome sequencing through chromosome sorting. PNAS *108*, 12–17.

Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W., and Arnheim, N. (1992). Whole genome amplification from a single cell: implications for genetic analysis. PNAS *89*, 5847–5851.

Chapter 8

Digital Whole Genome Amplification

8.1 Motivations and Summary

With the rapid development of next-generation sequencing techniques (Shendure and Ji, 2008; Metzker, 2010), genetic analysis has been routinely done genome-wide with single nucleotide resolution (Dewey et al., 2012). However, several challenges remained. First, when the analyzed samples are heterogeneous in their genetic content, especially when the frequency of the genetic variants is low, it is difficult to distinguish these variants in the background of sequencing and mapping artifacts (Devonec et al., 1990; Bhatia et al., 2012). Second, when the sample is very rare, such as in certain forensic and archeological applications (Thomas, 1993; Hanson and Ballantyne, 2005), as well as medical applications such as monitoring circulating tumor cells (Cristofanilli et al., 2004; Paterlini-Brechot and

Benali, 2007) and performing preimplantation screening (Mastenbroek et al., 2007; Harper et al., 2008), it is often difficult to obtain enough genetic materials for a comprehensive analysis of the whole genome.

To account for these challenges in whole genome analysis, single cell genome amplification methods were developed (Dean et al., 2002; Lao et al., 2008) and were demonstrated in analyzing tumor (Navin et al., 2011), sperm (Wang et al., 2012) and microbes (Yoon et al., 2011) etc. In Chapter 6, we introduced a method MALBAC with significantly improved amplification evenness, which allowed us to analyzed copy number variations (CNVs) and single nucleotide variations (SNVs) of single cells. The problem is, however, significant amplification bias and randomness still exist compared to the bulk samples that are not amplified, which makes it difficult for copy number variation analyses when the variant size is small (e.g. several kilobases). The amplification error generated in the first several cycles sometimes renders results that are not distinguishable with true variants.

Another challenge in single cell genomics as well as genetic analysis in general is the genome phase problem. The genome of a typical somatic human cell is diploid, which means there are two copies of very similar chromosome sequences each inherited from one of the parents (Alberts et al., 2007). It is therefore challenging to separate these two chromosomes in genetic analysis when using large amount of starting materials. There are several methods in the literature now on phasing individual genomes, however, they often required time-consuming

sample preparations such as cloning (Zhang et al., 2006; Levy et al., 2007; Kitzman et al., 2011; Suk et al., 2011), or complex instrumentation or devices to separate individual chromosomes in samples that are active in mitosis (Ma et al., 2010; Fan et al., 2011; Yang et al., 2011). It is therefore not very likely these methods are scalable and can be a general tool for whole genome analysis in general.

In this Chapter, we attempt to solve the problems mentioned above by performing digital whole genome amplification. By doing some preliminary experiments and with the preliminary data we have got from the sequencing runs, I hope to convey a message that this method could potentially change the way single cell genome is analyzed. Being totally independent of the amplification bias and unevenness, digital whole genome amplification (dWGA) can potentially reveals genome instabilities such as copy number variations at a much smaller scale. dWGA also enables whole genome phasing with minimal instrumentation requirement, which can be in principle done even in your kitchen.

8.2 The Genome Phase: an Introduction

One central goal of modern human genetics is to characterize the genetic variations of individuals, and to study their relations with various human phenotypes (Morley et al., 2004; Consortium, 2010). Human gene structure is complex, often contains multiple coding and regulatory regions, and whether the genetic variations are associated on the same chromosome (in cis), or on opposite homolog chromosomes (in trans) have different phenotypic

consequences. One example is the well-established phenomenon of cis and trans compound heterzygosity (Van Driest et al., 2004; Ingles et al., 2005), where the two heterozygous non-synonymous mutations in a certain gene can result in either one normal version (as in cis) or no normal version (as in trans) of the gene, as is shown in Figure 8.1.

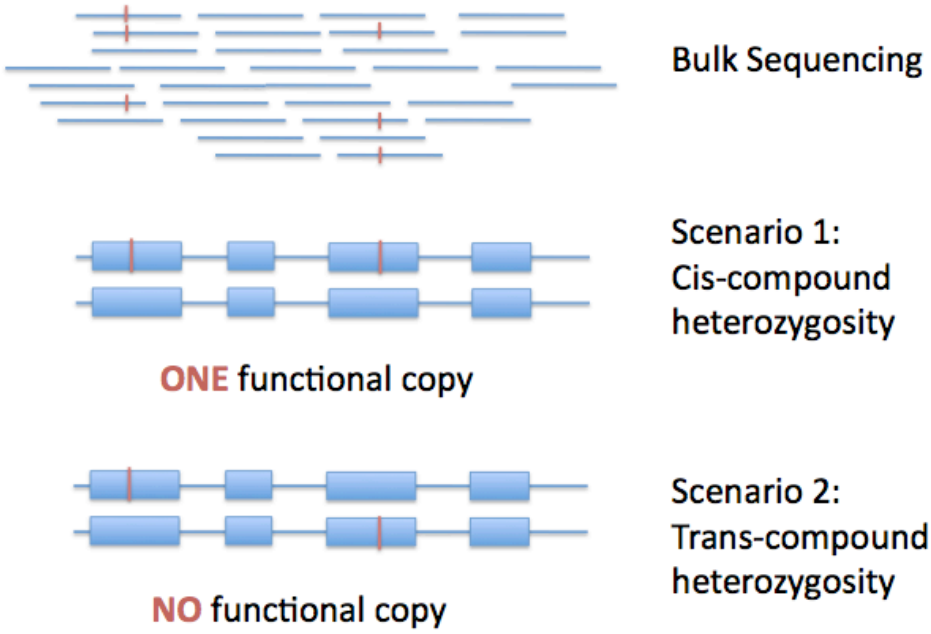


Figure 8.1: The importance of the phase information in the correct interpretation of human genome. The phase information is often loss in bulk genome analysis. Here we list two different scenarios in which the phenotypic consequences are completely different, but they are often indistinguishable in sequencing the bulk genome.

Such phenomenon is further complicated by the complex structure of human genes. The length of the functional unit of a human gene ranges from ~10kb to ~1Mb. These units often contain multiple exon fragments, variable promoter sequences and enhancer/insulators

(Watson et al., 2007; Krebs et al., 2009). The human genome has about one single nucleotide variant per one kilobase (Altshuler et al., 2010; Consortium, 2010), which corresponds to about 10 to 1000 variants per functional unit. These variants as a whole determine the function of the gene (Montgomery et al., 2010; Pickrell et al., 2010). Allele-specific expression has been found common in human genes and related to diseases such as cancer and neurological disorders (Chen et al., 2008; Palacios et al., 2009; Chamberlain and Lalande, 2010), and recent studies have shown that cis-acting sequence variations significantly influences methylation patterns and gene expression (McDaniell et al., 2010; Zhang et al., 2010). A complete understanding of these phenomena and their consequences in human variations and diseases cannot be obtained without resolving the combination of alleles at different loci on the same chromosome, that is, the haplotypes (Consortium, 2005; Frazer et al., 2007). Without the haplotype information, the description of individual human genome is incomplete and the interpretation is error-prone.

Population-level haplotype information can be inferred from pedigree studies and linkage disequilibrium data (Ardlie et al., 2002; Roach et al., 2010). However, these population based studies often fail to predict the haplotype structure of rare, individual variants or regions with high recombination rate. Determination of the haploid structure of an individual at a whole genome scale is challenging and currently lacking cost-effective approaches. Mate-pair Sanger sequencing was first used for phasing an individual's whole genome (Levy et al., 2007), but it is too costly and labor-intensive. Fosmid pool-based genome sequencing was

demonstrated in whole genome haplotyping (Kitzman et al., 2011; Suk et al., 2011), but is limited by the time-consuming and costly cloning steps and requires large amount of input materials. Haplotype analysis has also been demonstrated by physically separating single human chromosomes before amplification and genotyping (Ma et al., 2010; Fan et al., 2011; Yang et al., 2011). However, besides the extra instrumentation complexity and being time consuming, these methods require cells in metaphase, and are therefore not suitable for specimens that are either not active in mitosis or not culturable in labs, such as frozen human tissues. We note that a recent report described a method to perform haplotyping by diluting DNA fragments before whole genome amplifying and sequencing each fragment (Peters et al., 2012). However, this study required ~10-20 cells with high sequencing depth (~100x in total).

8.3 Genome Phasing by Digital Whole Genome Amplification (dWGA)

Here we report a method named digital Whole Genome Amplification (dWGA) for whole genome sequencing single somatic human cells with resolved haploid structure. dWGA uses minimal input material (1-3 single cells), does not require live cells, cell culturing and complicated instruments and devices for separating chromosomes, and does not significantly increase the cost for whole genome sequencing. We also note that performing single cell analysis by dWGA can in principle reveal cell-to-cell variations with haploid resolution in genetically heterogeneous specimens, such as in cancer (Bhatia et al., 2012).

Here we describe the procedure of dWGA. We obtained isolated single cells by several different methods for different samples. For the blood sample from an anomalous individual, we first lysed the cellular membrane by mild detergent treatment. After the treatment, only the nuclei of the nucleated cells are visible under the microscope. We then used mouth pipetting to isolate each single nucleus into a separate reaction well. For culture cells such as a cancer cell line SW480 and a human B-lymphoblast cell line GM12878, we stained the DNA with a living cell stain Vybrant Ruby and used a flow cytometer to select for the diploid population.

The isolated single cells were then lysed with protease treatment and were separated into multiple reaction wells simply by pipetting, as is shown in Figure 8.2. The DNA molecules (~150 kb after lysis) from each well were separately amplified by a recently developed whole genome amplification method MALBAC (Multiple Annealing and Looping Based Amplification Cycles) (Chapter 6) and were sequenced with barcodes on a Illumina HiSeq 2000 platform. Then we mapped the sequences from different reaction wells separately to the human reference genome (Figure 8.2).

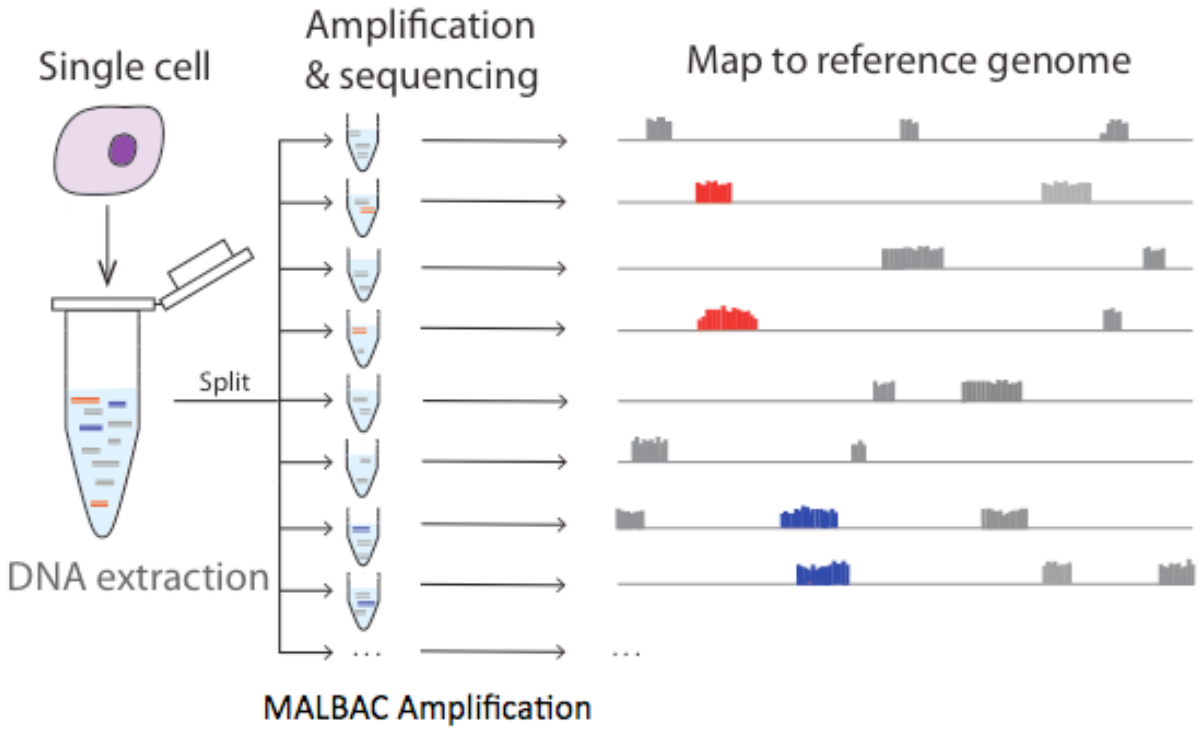


Figure 8.2: The procedure of dWGA using MALBAC amplification. The DNA molecules from single cells are individually barcoded, sequenced and mapped onto the human genome. Shown here in colors (blue and red), are the homologue chromosome fragments that are separated into different reaction wells, they represent different haplotype contribution of the paternal and maternal heritages.

By dividing single cell genome into multiple portions before DNA amplification, the two homologous DNA molecules are statistically separated. The chance of the homologous DNA molecules not separated into different wells is 1/N, with N being the number of wells used in the experiment. Therefore, genetic features identified (such as single nucleotide variants (SNVs) and small insertions and deletions (indels)) within each 'amplification blocks' are

homozygous and are linked on the same chromosome.

He we show the preliminary sequencing result in Figure 8.3. After sequencing and mapping the DNA molecules from the blood samples from an individual, we obtained sequencing reads covering four heterozygous SNVs G/C, A/C, A/G, T/C, but we do not know which 4 SNVs are associated on the same chromosome. From sequencing the 'single fragment amplified' genomes in two reaction wells and identify the SNVs in each well, we can safely infer that the four SNVs CCAT are associated and GAGC are associated on the opposite chromosome.

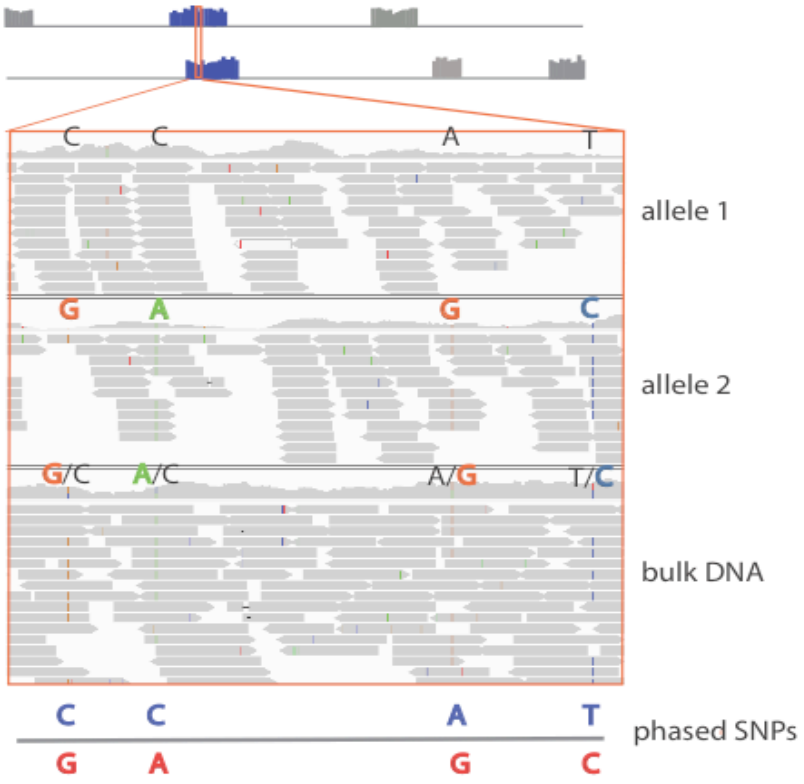


Figure 8.3: An example showing the principle of using dWGA to phase the SNPs identified in bulk genome sequencing.

By doing this, we obtained haplotype blocks of ~100kb. For getting longer haplotype blocks, we 'stitched' the amplification blocks into much longer haplotype blocks (~1Mb) by comparing other reaction wells from the same cell, as is shown in Figure 8.4, the experimental data is shown in Figure 8.5.

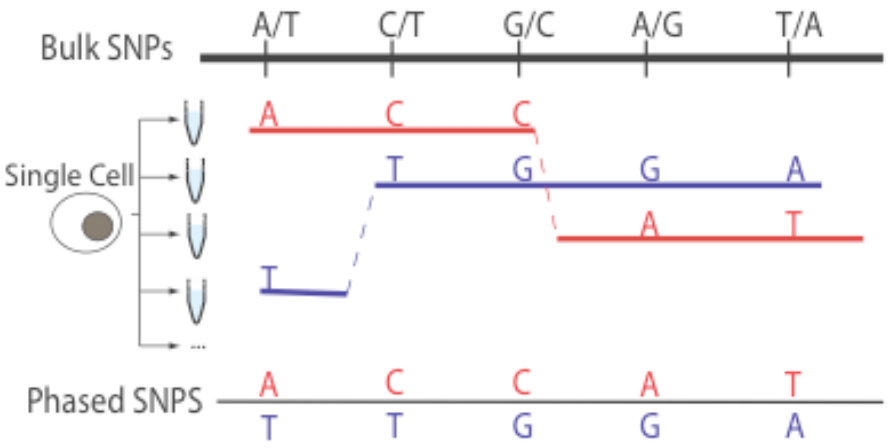


Figure 8.4: The procedure of comparing all reaction wells from a single cell to reconstitute a much larger haplotype block

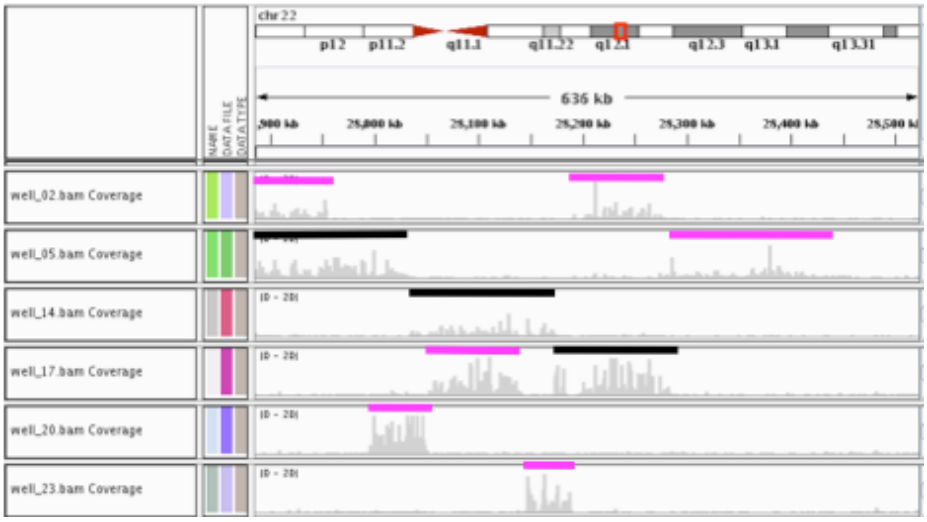


Figure 8.5: The experimental data showing how small fragments can be 'stitched' together to form much bigger fragments by sequencing a single cell.

The Haplotype block size can be further increased by comparing the resolved haplotypes between different cells, as in shown in Figure 8.6.

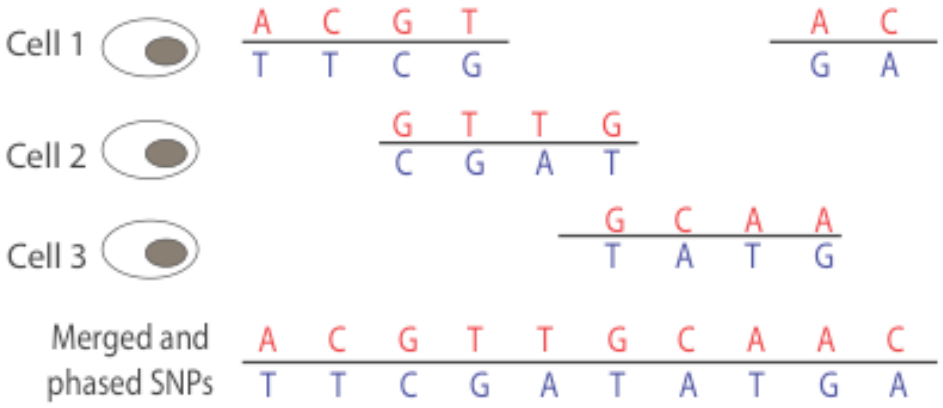


Figure 8.6: By comparing the resolved haplotype blocks from about 3-4 cells, the haplotype blocks can be further extended to several megabases.

To verify the phasing results by dWGA on single cells, we lightly sequenced the genome from the parents of the anonymous donor (Figure 8.7). Genome phase information can be inferred using the family trio information and we confirmed they are consistent with the result we obtained from dWGA.

We note that the sample preparation before sequencing for dWGA can be completely in <3 days by a single technician at a cost of ~$500, which is significantly lower than the cost of whole genome sequencing. The procedure also does not require any complicated devices or instrumentation, which can be done in principle anywhere in the world.

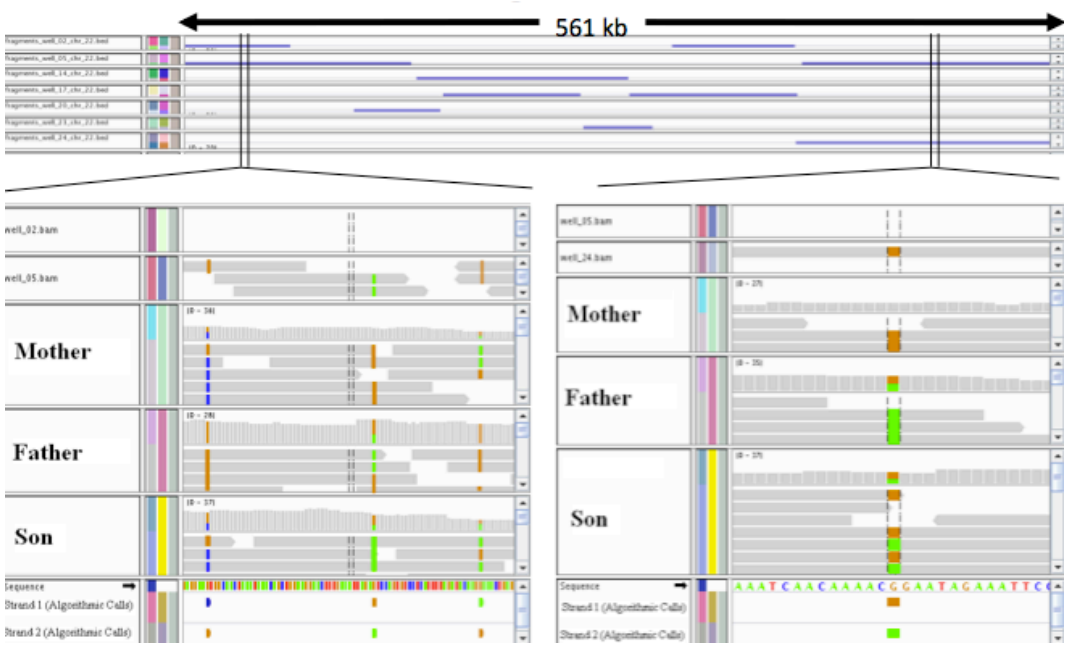


Figure 8.7: Phasing using the family trio information, the results compare well with phasing using dWGA.

8.4 Digital Counting of Copy Number Variations (CNVs) by dWGA

Now we explore another aspect of dWGA. As we discussed in Chapter 6, when the sample for genome analysis is heterogeneous or rare, single cell genome amplification and sequencing is necessary to comprehensively analyze the genome. However, even with the improved method MALBAC, there are still substantial amplification variations compared with an unamplified sample, which may complicate the analysis such as CNV profiling. This is less of a problem if the CNV spans a big genome fragment (several megabases to a whole chromosome), because the sequencing reads can be binned into huge statistic window at the price of sacrificing resolution. However, most newly generated CNVs are small (Hussein et al., 2011) and may be masked by the amplification noise from single cell WGA.

Being independent of the amplification noise, dWGA counts the absolute number of the genome fragments. Assume we dilute the single cell genome into N number of reaction wells, assuming N is large, then a three-copy region should shows sequencing reads in three of the reaction wells. Although the three copies may be amplified differently due to amplification noise, it does not affect counting digitally the number of copies (Figure 8.8).

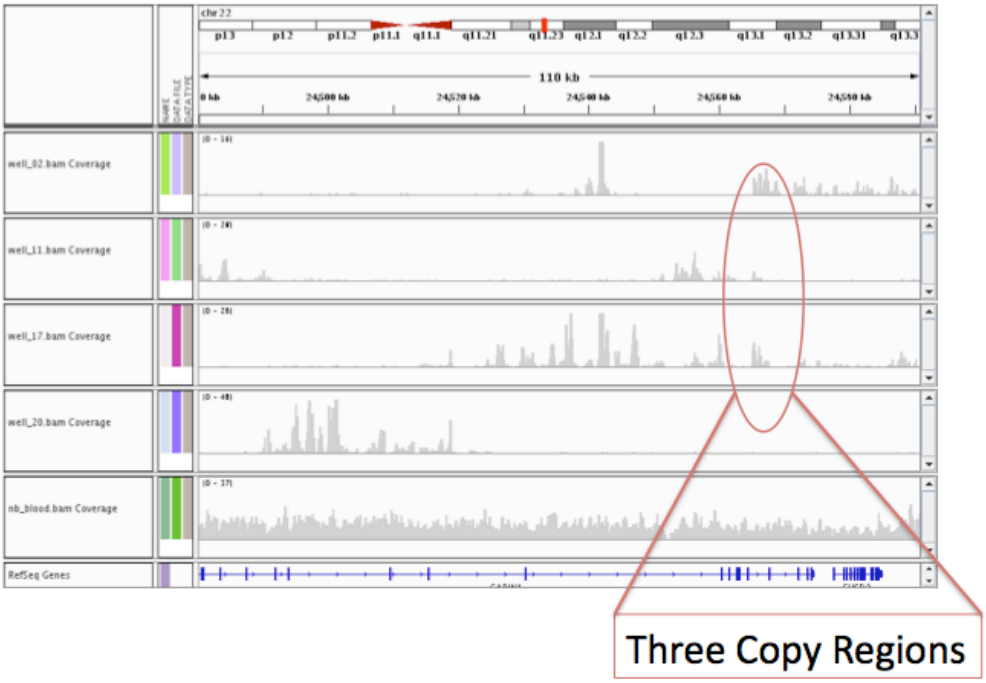


Figure 8.8: A three-copy region of the genome reflected by the fact that sequencing reads are mapped to such region in three reaction wells.

As a summary, dWGA separates homologous chromosome and enables a comprehensive analysis of single cell genome with resolved haplotype structure. dWGA also bypasses problem of unevenness amplification in single cell genome analysis, enabling high-resolution CNV analysis in a single cell.

References:

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). Molecular Biology of the Cell (Garland Science).

Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

Ardlie, K.G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics *3*, 299–309.

Bhatia, S., Frangioni, J.V., Hoffman, R.M., Iafrate, A.J., and Polyak, K. (2012). The challenges posed by cancer heterogeneity. Nature Biotechnology *30*, 604–610.

Chamberlain, S.J., and Lalande, M. (2010). Neurodevelopmental disorders involving genomic imprinting at human chromosome 15q11–q13. Neurobiology of Disease *39*, 13–20.

Chen, X., Weaver, J., Bove, B.A., Vanderveer, L.A., Weil, S.C., Miron, A., Daly, M.B., and Godwin, A.K. (2008). Allelic imbalance in BRCA1 and BRCA2 gene expression is associated with an increased breast cancer risk. Hum. Mol. Genet. *17*, 1336–1348.

Consortium, T.1000 G.P. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Consortium, T.I.H. (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

Cristofanilli, M., Budd, G.T., Ellis, M.J., Stopeck, A., Matera, J., Miller, M.C., Reuben, J.M., Doyle, G.V., Allard, W.J., Terstappen, L.W.M.M., et al. (2004). Circulating tumor cells, disease progression, and survival in metastatic breast cancer. New England Journal of Medicine *351*, 781–791.

Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. PNAS *99*, 5261–5266.

Devonec, M., Bringuier, P.P., Hijazi, A., Dutrieux-Berger, N., Revillard, J.P., and Perrin, P. (1990). Heterogeneity of DNA index in renal-cell carcinoma. Prog. Clin. Biol. Res. *348*, 35–48.

Dewey, F.E., Pan, S., Wheeler, M.T., Quake, S.R., and Ashley, E.A. (2012). DNA sequencing: clinical applications of new DNA sequencing technologies. Circulation *125*, 931–

944.

Van Driest, S.L., Vasile, V.C., Ommen, S.R., Will, M.L., Tajik, A.J., Gersh, B.J., and Ackerman, M.J. (2004). Myosin binding protein C mutations and compound heterozygosity in hypertrophic cardiomyopathy. Journal of the American College of Cardiology *44*, 1903–1910.

Fan, H.C., Wang, J., Potanina, A., and Quake, S.R. (2011). Whole-genome molecular haplotyping of single cells. Nature Biotechnology *29*, 51–57.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Hanson, E.K., and Ballantyne, J. (2005). Whole genome amplification strategy for forensic genetic analysis using single or few cell equivalents of genomic DNA. Anal. Biochem. *346*, 246–257.

Harper, J., Sermon, K., Geraedts, J., Vesela, K., Harton, G., Thornhill, A., Pehlivan, T., Fiorentino, F., SenGupta, S., Die-Smulders, C. de, et al. (2008). What next for preimplantation genetic screening? Hum. Reprod. *23*, 478–480.

Hussein, S.M., Batada, N.N., Vuoristo, S., Ching, R.W., Autio, R., Närvä, E., Ng, S., Sourour, M., Hämäläinen, R., Olsson, C., et al. (2011). Copy number variation and selection during reprogramming to pluripotency. Nature *471*, 58–62.

Ingles, J., Doolan, A., Chiu, C., Seidman, J., Seidman, C., and Semsarian, C. (2005). Compound and double mutations in patients with hypertrophic cardiomyopathy: implications for genetic testing and counselling. Journal of Medical Genetics *42*, e59–e59.

Kitzman, J.O., Mackenzie, A.P., Adey, A., Hiatt, J.B., Patwardhan, R.P., Sudmant, P.H., Ng, S.B., Alkan, C., Qiu, R., Eichler, E.E., et al. (2011). Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat. Biotechnol. *29*, 59–63.

Krebs, J.E., Goldstein, E.S., and Kilpatrick, S.T. (2009). Lewin's Genes X (Jones & Bartlett Publishers).

Lao, K., Xu, N.L., and Straus, N.A. (2008). Whole genome amplification using single-primer PCR. Biotechnol J *3*, 378–382.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The Diploid Genome Sequence of an Individual Human. PLoS Biol *5*, e254.

Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K., and Song, Q. (2010). Direct determination of molecular haplotypes by chromosome microdissection. Nature Methods *7*, 299–301.

Mastenbroek, S., Twisk, M., van Echten-Arends, J., Sikkema-Raddatz, B., Korevaar, J.C., Verhoeve, H.R., Vogel, N.E.A., Arts, E.G.J.M., De Vries, J.W.A., Bossuyt, P.M., et al. (2007). In vitro fertilization with preimplantation genetic screening. New England Journal of Medicine *357*, 9–17.

McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., et al. (2010). Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. Science *328*, 235–239.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nature Reviews Genetics *11*, 31–46.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773–777.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S., and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. Nature *430*, 743–747.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Palacios, R., Gazave, E., Goñi, J., Piedrafita, G., Fernando, O., Navarro, A., and Villoslada, P. (2009). Allele-Specific Gene Expression Is Widespread Across the Genome and Biological Processes. PLoS ONE *4*, e4150.

Paterlini-Brechot, P., and Benali, N.L. (2007). Circulating tumor cells (CTC) detection: clinical impact and future directions. Cancer Letters *253*, 180–204.

Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., et al. (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature *487*, 190–195.

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

Roach, J.C., Glusman, G., Smit, A.F.A., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Science *328*, 636–639.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology *26*, 1135–1145.

Suk, E.-K., McEwen, G.K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D.T., McLaughlin, S., Peckham, H., et al. (2011). A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res *21*, 1672–1685.

Thomas, K.D. (1993). Molecular biology and archaeology: A prospectus for inter-disciplinary research. World Archaeology *25*, 1–17.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. Cell *150*, 402–412.

Watson, J.D., Baker, T.A., Bell, S.P., Gann, A., Levine, M., Losick, R., and CSHLP, I. (2007). Molecular Biology of the Gene (Benjamin Cummings).

Yang, H., Chen, X., and Wong, W.H. (2011). Completely phased genome sequencing through chromosome sorting. PNAS *108*, 12–17.

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C., Duffy, S., and Bhattacharya, D. (2011). Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists. Science *332*, 714–717.

Zhang, D., Cheng, L., Badner, J.A., Chen, C., Chen, Q., Luo, W., Craig, D.W., Redman, M., Gershon, E.S., and Liu, C. (2010). Genetic control of individual differences in gene-specific methylation in human brain. Am. J. Hum. Genet. *86*, 411–419.

Zhang, K., Zhu, J., Shendure, J., Porreca, G.J., Aach, J.D., Mitra, R.D., and Church, G.M. (2006). Long-range polony haplotyping of individual human chromosome molecules. Nat. Genet. *38*, 382–387.

Chapter 9

Single Cell Transcriptome and Genome Accessibility Studies

9.1 Motivations

Most organisms originate from a single cell. However, instead of simply replicating itself, the single cell develops into a complex organism that often consists of different cell types (Keller et al., 2008; Gilbert, 2010). It is the interactions of these descendent cells that provide normal function for the organism. There are thousands of different cell types in a single human being. They have very different appearances and functions, but they all come from a single cell and have the same (or very similar) genome. To study how identical genome generates functional differences between cells requires pushing single cell techniques beyond genome analysis tan(Tang et al., 2009). In this chapter, we introduce our recent efforts in transcriptome

profiling of single cells. We further ask why different cells have different transcriptome and how they are regulated, and these questions prompted us to further study the genome accessibility (Bell et al., 2011), at the single cell level. The work described here are ongoing projects; we are in the process of performing sequencing and data analysis. Here we present preliminary data as a proof-of-principle demonstration of our efforts on studies of transcriptome and genome accessibility of single cells.

9.2 Genome Instability and Consequences

Our effort in studying single cell transcriptome started when we observed that the single cancer cells we have studied often exhibit different genome with other cancer cells (Yachida et al., 2010; Navin et al., 2011)(Chapter 6). Actually, genome instability is not a specific problem to cancer. It has been reported in almost every different cell types, such as in embryonic stem cells (Lefort et al., 2008, 2009), neurons (Rehen et al., 2005; Yurov et al., 2007), gamete cells (Downie et al., 1997; Wang et al., 2012). It was found that identical twins and different tissues (Hall, 1988; Youssoufian and Pyeritz, 2002) often have copy number variations in certain parts of their genomes, indicating that genome instability does exist very early in development.

One key question, however, is whether these genome changes really have functional consequences, and if so, how does the genome instability shape the transcriptome or the proteome of different cells (Williams et al., 2008; Habermann et al., 2011). It was known that

tumor originates from one or few cells with a set of key mutations that disrupt the normal cellular function (Hanahan and Weinberg, 2011). However, associated with these 'driver' mutations, there are other 'passenger' mutations with insignificant functions, and it is often difficult to distinguish them (Greenman et al., 2007; Akavia et al., 2010). Recently, it was found that in human embryonic stem cells (Lefort et al., 2008) and induced pluripotent cells (Gore et al., 2011; Hussein et al., 2011) , there are multiple point mutations and copy number variation generated in the early cycles before being selected out in the subsequent passages, which indicates these genome variations change the viability and thus the function of the cells. Since mutations happen in the unit of single cells, it is important to capture both the genome and the transcriptome of the same cell and study their correlations.

9.3 Genome and Transcriptome of the Same Cell

It seems technically impossible to get both the transcriptome and the genome from a single cell, because it is extremely difficult to separate DNA and RNA molecules based on their chemical and physical properties (Metzenberg, 2007; Nelson and Cox, 2008). Luckily, mother nature provides a natural solution to the problem. In mammalian cells, DNA and RNA molecules have very different distribution pattern. Genome DNA molecules are exclusively confined in the cellular nuclei and matured mRNA molecules preferentially locate in the cytoplasm. We used this to first separate DNA and mRNA molecules from a single cell, as is shown in Figure 9.1.

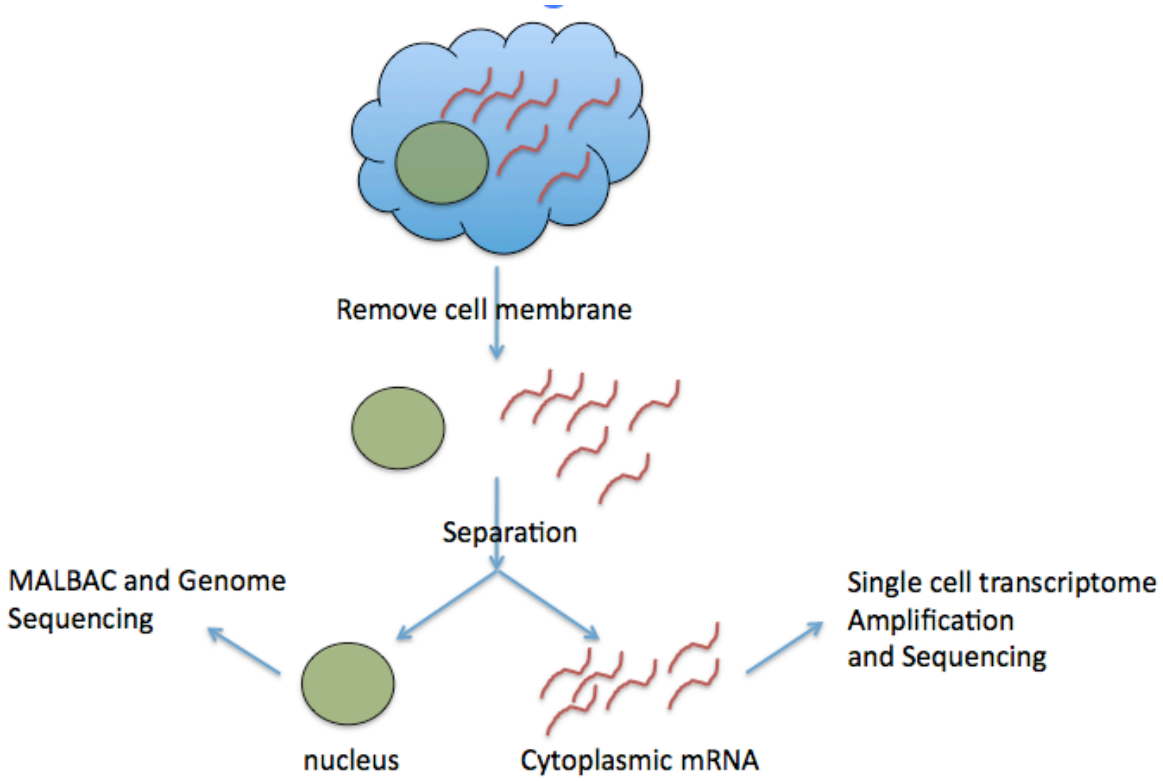


Figure 9.1 The procedure of obtaining both the genome and the transcriptome of a single cell

After isolating a single cell by mouth pipetting, flow cytometry, or laser dissection, we treated the cell with low percentage of mild detergents with controlled duration of time to remove the cellular membrane. The nuclear membrane remains intact because of the protection of the much denser protein structure in the nuclear envelope (2010).

We then centrifuged the reaction tube in high speed to separate the nucleus with the cytoplasm based on the difference of density. We then used the nucleus to do whole genome amplification and sequencing using the method MALBAC introduced in Chapter 6. And we

use the supernatant containing most cytoplasmic RNA to perform transcriptome analysis on the same cell.

Single cell transcription analysis has been routinely used to study the cellular heterogeneity in gene expression on specific genes (Chiang and Melton, 2003; Tietjen et al., 2003; Dalerba et al., 2011). With the rapid developments of the high throughput genotyping methods such as microarray (Heller, 2002) and next generation sequencing (Metzker, 2010), single cell transcription analysis is becoming high-throughput and the whole transcriptome can be analyzed at the single cell level.

Several methods have been reported on sequencing the whole transcriptome on single cells (Kurimoto et al., 2006; Tang et al., 2009; Islam et al., 2011). We first try using the method described in (Tang et al., 2009) to perform whole transcriptome analysis on the isolated cytoplasm of the single cells. In brief, the method contains five steps: First, the poly-A (+) mRNA molecules are reverse transcribed using a primer containing a fixed common sequences (UP1) at its 5' end and a poly-T at its 3' end. Then after removing the free primers, we treated the first strand cDNA with terminal transferase and adenine. Then a second primer with another common sequences (UP2) at 5'end and a poly-T at 3'end extends the second strand cDNA. After doing this, we have a double-strand cDNA generated from the original RNA template, each end with a common sequence to be further amplified by PCR to enough amount for whole transcriptome sequencing.

We sequenced both the genome and the transcriptome of six cells from a cancer cell line SW480 (ATCC), and we aimed to look at the correlation between them. For example, it is known in Drosophila, dosage compensation maintains the transcript quantity between male and females through epigenetic regulation (Baker et al., 1994). Similar situations happen on the X chromosome in humans, in which one of the X chromosomes are imprinted to maintain transcription balance in females, known as X- chromosome inactivation (Avner et al., 2001). We want to look at whether such compensation effects exist in autosomes, and how the cellular transcriptome is going to respond to an extra copy of certain chromosome fragments.

We sequenced the genomes with low coverage aiming for analyzing copy number variations, and we sequenced the transcriptome at ~100x trying to detect the variations of transcriptome between cells (Morozova et al., 2009). However, we failed to see the correlation of the copy number in the genome and the transcriptome profile of the same cell (Figure 9.2).

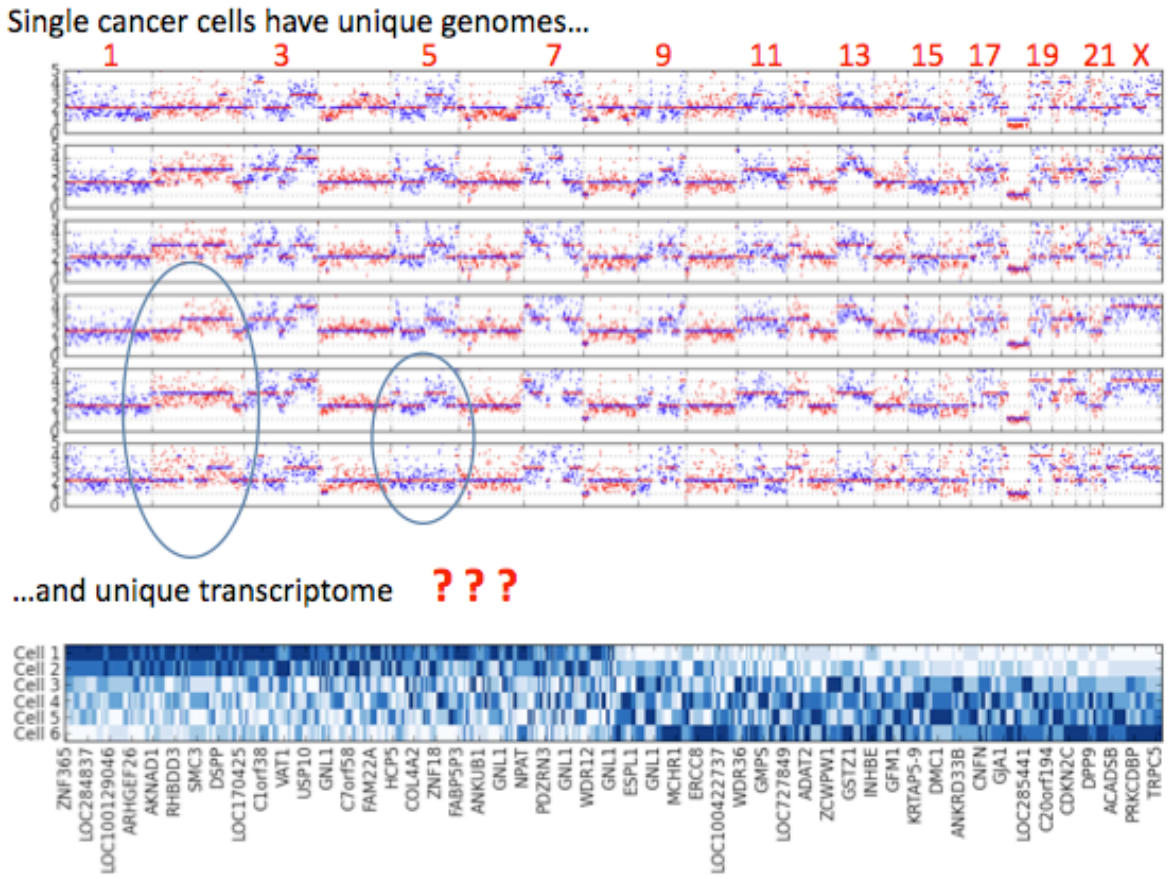


Figure 9.2 Sequencing both the genome and the transcriptome of six single cells. Shown here are the copy number profiles of different chromosomes in the human genome. And we plot the mappable read number of a list of genes on the transcriptome, with darker color representing higher depths of sequencing reads.

Such result could indicate two possibilities: One, the transcriptome is extremely noisy for cancers, different cells at a certain time point are completely different regardless of their DNA contents (Blake et al., 2003; Taniguchi et al., 2010). Or it could mean the method for whole transcriptome analysis is not valid, or not good enough to pick up such differences.

To verify whether the 'noise' is biological or technical. We spiked in synthesized mRNA molecules in the single cells and perform whole transcriptome amplification and sequencing. We sequenced six technical replicates to characterize the technical variations due to single cell transcriptome amplification and sequencing.

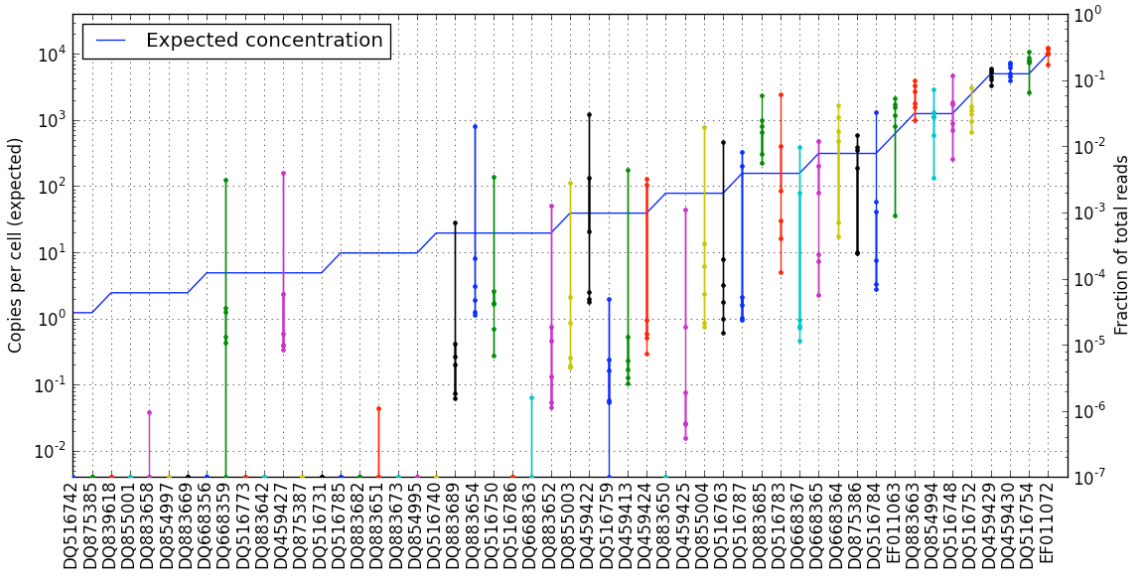


Figure 9.3 Sequencing of the spike-in mRNA molecules. The blue curve represents the expected sequencing coverage from each molecule, and the colored bars connected six technical replicates.

For mRNA molecules with more than a thousand copies, the variation of the six replicates is less than 10 fold. However, at lower copy numbers (tens of copies), the technical variations span ~ 3 orders of magnitude. It is therefore not very likely we can see the transcriptional differences due to copy number variations because of the technical noise generated in mRNA amplification.

9.4 Analyzing the Transcriptome of Single Cells by MALBAC

In Chapter 6, we discussed a method named MALBAC aiming to amplify and sequence the whole genome of a single cell. Here we slightly modify this method for amplification and sequencing the transcriptome of a single cell.

There are two main challenges we have to deal with when we try to amplify very small amount of genetic materials, such as a single cell. First is the efficiency of amplification. In the previously mentioned single cell transcriptome method (Tang et al., 2009), there are five reaction steps before a full amplicon representing a single mRNA transcript can be made, which harms the efficiency. The second consideration is the amplification 'noise', meaning given a full amplicon, what is the variation of total concentration after being amplified to an enough amount for downstream genetic analysis such as high throughput sequencing. PCR is known for robust amplification of hundreds of molecules (Saiki et al., 1988). However, when the starting amount is down to several copies, the variations of the total molecular concentration after amplification can be huge due to the kinetic variations in the first several PCR cycles (Shiroguchi et al., 2012).

The unique properties of MALBAC solve both of these problems. First, we simplified the reaction to only two steps: A reverse transcription step that converts the mRNA templates to cDNA molecules, and a MALBAC step to amplify all the cDNA molecules that are converted. One potential caveat here is the contamination from the genome DNA, and we

confirmed that a simple centrifugation step successfully separate the DNA and the cytoplasmic mRNA in >90% of the cells. Second, MALBAC provides a close-to-linear pre-amplification step, which creates about 100 molecules per input template before exponential amplification. Therefore the amplification noise is expected to be significantly reduced.

We amplified five single-cell level mRNA replicates, together with the spike-in internal controls for ten cells. Five of them we used MALBAC (red +), and the other half by the reported single cell transcriptome method (black dots) (Tang et al., 2009).

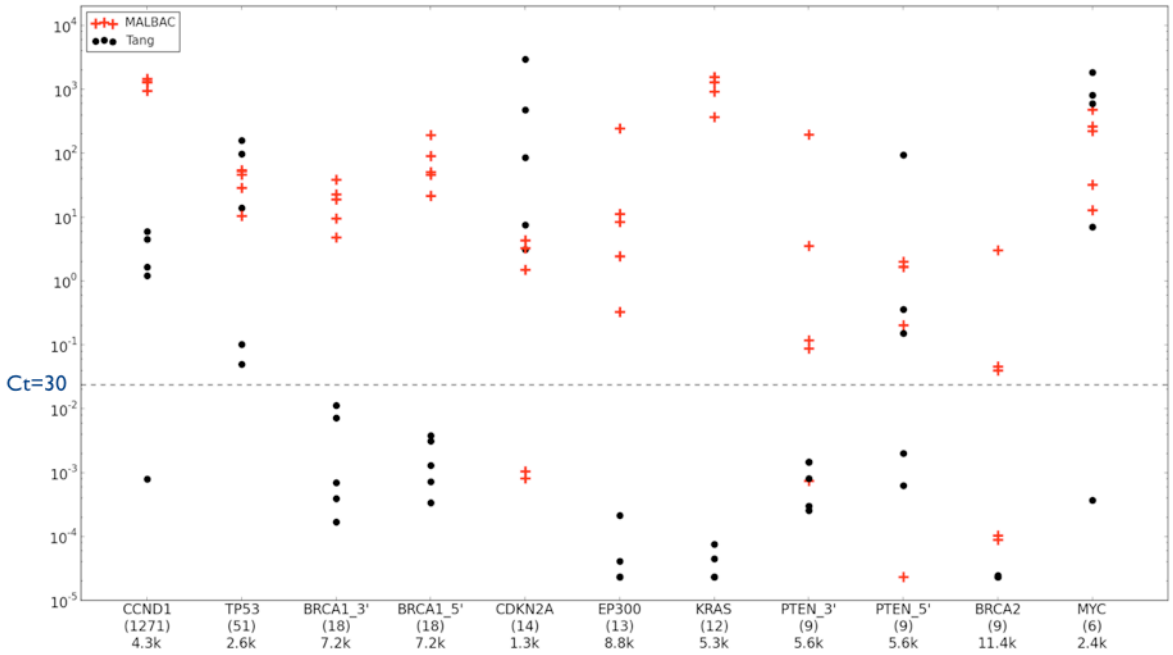


Figure 9.4 Comparison of the transcriptome amplification evenness using MALBAC and a reported amplification method (Tang et al., 2009). x-axis shows different genes with their estimated copy number and length. Y axis shows the relative amount estimated by qPCR. We set a threshold at Ct=30, points below this threshold are considered not amplified because of the rare amount in the total amplified products.

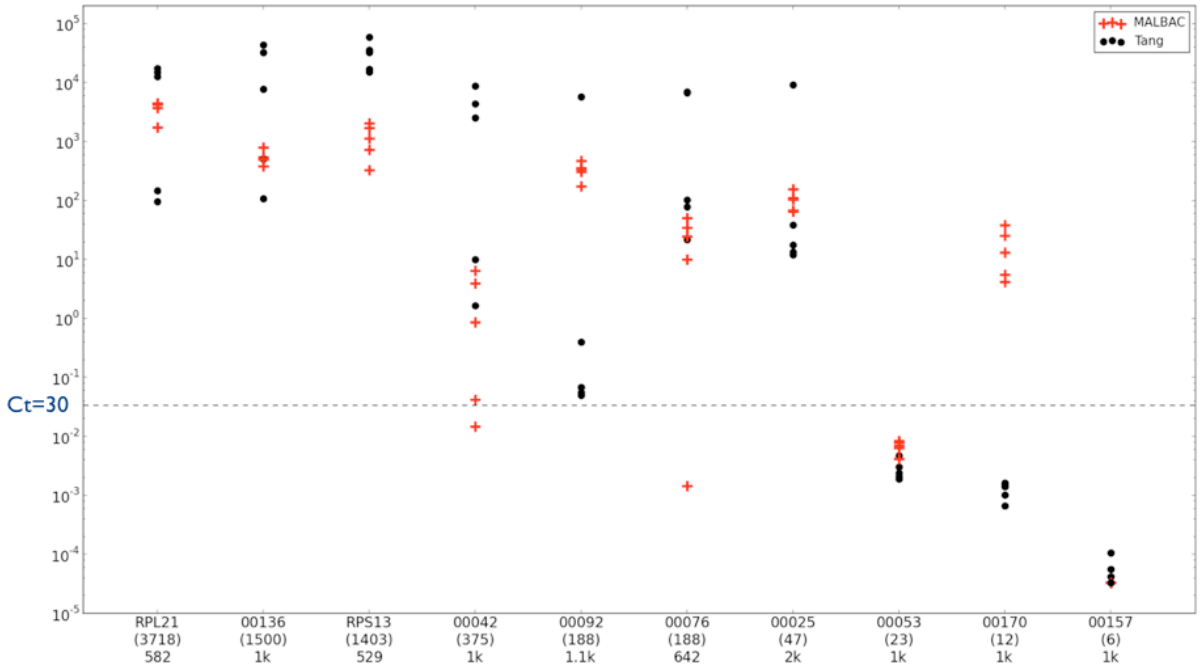


Figure 9.5 Comparison of the transcriptome amplification evenness using MALBAC and a reported amplification method (Tang et al., 2009). x-axis shows different spike-in mRNA molecules with their estimated copy numbers and lengths. Y axis shows the relative amount estimated by qPCR. We set a threshold at Ct=30, points below this threshold are considered not amplified.

Compared with the previous method (Tang et al., 2009), MALBAC is significantly better in both sensitivity (the percentage of transcripts that are amplified with Ct<30) and variability (how close the 'dots' are). In Figure 9.6, we compare the sensitivity (right) and technical variations (left) by the two methods.

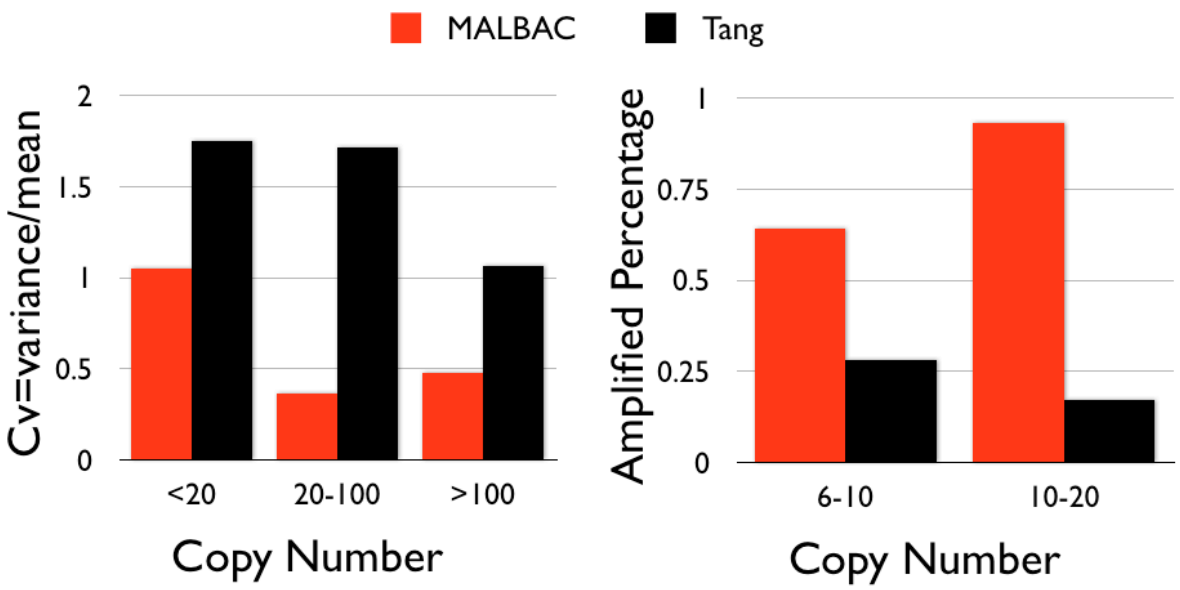


Figure 9.6 Comparison of the transcriptome amplification efficiency and evenness using MALBAC and a reported amplification method (Tang et al., 2009). x-axis shows the copy number of the genes under analysis. (Left) Y axis shows the coefficient of variance of the two amplification methods; (Right) Y axis shows the percentage of genes that are amplified by the two methods.

Interesting, we observed a correlation of gene length and amplification efficiency by either of the two methods. The Tang method selectively amplified shorter genes, and MALBAC amplified longer genes (>2.5 kb) better than shorter genes. ~70% of the oncogenes and tumor suppressor genes are >2.5kb, to profile these genes in single cells, MALBAC provide superb sensitivity and low variability.

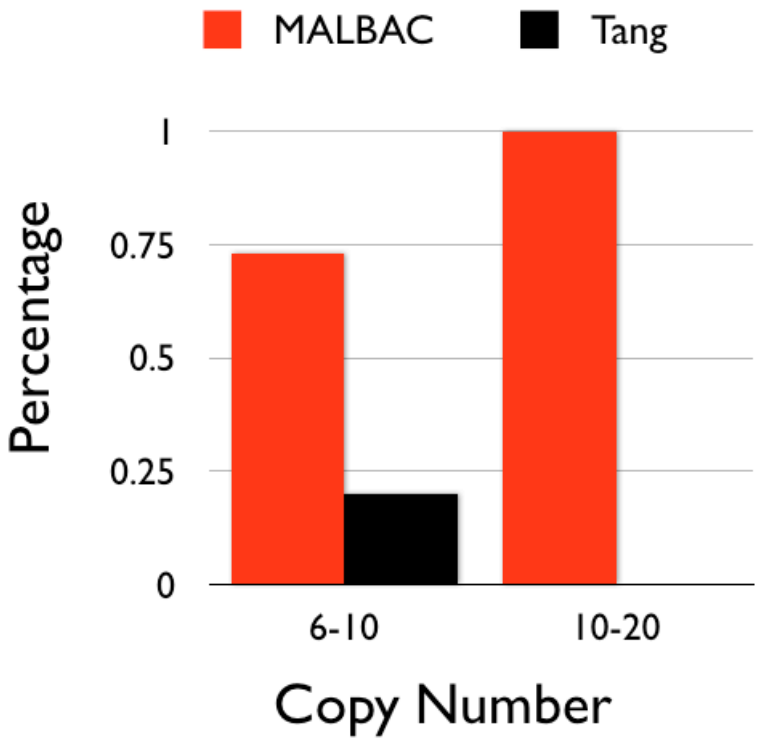


Figure 9.7 Comparison of the transcriptome amplification efficiency for genes longer than 2.5kb using MALBAC and a reported amplification method (Tang et al., 2009). x-axis shows the copy number of the genes under analysis. Y axis shows the percentage of genes that are amplified by the two methods.

9.5 The Missing Link: Epigenetics at the Single Cell Level

One key question that is fundamental to biology is how variations at the genome level change the cellular functions. At the single cell level, although the tools are not perfect yet, we have a set of tools to study genome (Dean et al., 2001; Lasken, 2007; Chapter 6) and transcriptome variations (Klein et al., 2002; Tang et al., 2009; Dalerba et al., 2011). However, transcriptome variations are normally not directly caused by genome variations. Epigenetic factors such as

DNA methylation (Li et al., 1993; Bird, 2002) and histone modifications (Strahl et al., 2000; Berger, 2002) have significant effects on the transcriptome of a cell. For example, 5-methylcytosine (5-mC) replacements of the unmethylated cytosine in a gene promoter are often associated with silencing of the gene, which is often correlated with the silence histone marker such as H3K27me3 (Barski et al., 2007).

These Epigenetic factors, together with the genetic factors, shape the transcriptome, and thus the function of a cell, which are shown to be very dynamic in lots of important biological processes, such as early development of an embryo (Mikkelsen et al., 2007; Smith et al., 2012) and cancer metastasis (Szyf et al., 2004; Gupta et al., 2010). However, to our knowledge, there is no current method enables genome-wide epigenetic analysis on single cells. And we can never understand the functional correlation of genome and transcriptome without dealing with the epigenetic factors at the single cell level.

Single cell genome-wide epigenetic analysis is extremely challenging. Not only do we have to deal with the amplification unevenness problem in single cell amplification, we have to also figure out a way to 'encode' the epigenetic information into the 'sequencing reads' containing AGCT, which are the only output we can get from a typical sequencing run. In ensemble genome sequencing, the information of methylation can be extracted by bisulfite sequencing, in which the unmethylated cytosines are converted into uracils, which are read out as thymines, while the methylated cytosines remain unchanged (Frommer et al., 1992).

However, bisulfite conversion damages 90% of the DNA templates (Grunau et al., 2001), which is not a problem if we have a large amount of starting materials, but is deadly to single cell analysis. Indeed, we obtained less than 1% genome coverage after performing bisulfite conversion and MALBAC amplification on single cells. Similarly, chromatin immunoprecipitation sequencing (ChIP-seq), which is widely used to profile the genome-wide chromatin state, is not likely to work in single cells, because of the low efficiency of the immunoprecipitation process.

We therefore seek other route for single cell analysis. DNaseI sensitivity profiling has been a well-established essay to study the accessibility of chromatin structures (Sabo et al., 2006; Degner et al., 2012). The regions exhibit hypersensitivity of DNaseI are often promoters of highly expressed genes, indicating these sensitive sites can be also accessed by transcription initiating complex such as RNA polymerases and transcription factors (Boyle et al., 2008).

We then explore the possibility of getting the genome-wide chromatin accessibility information on single cells, by treating single cells with DNaseI before whole genome amplification using MALBAC. The idea is simple, those regions digested by the DNase should not show up in the sequencing reads, therefore we should be able to see less reads mapped to a certain region if such region is sensitive to DNaseI. We lightly sequenced five DNaseI treated cells with no DNaseI control, and we mapped the reads across the human genome, as is shown in Figure 9.7.

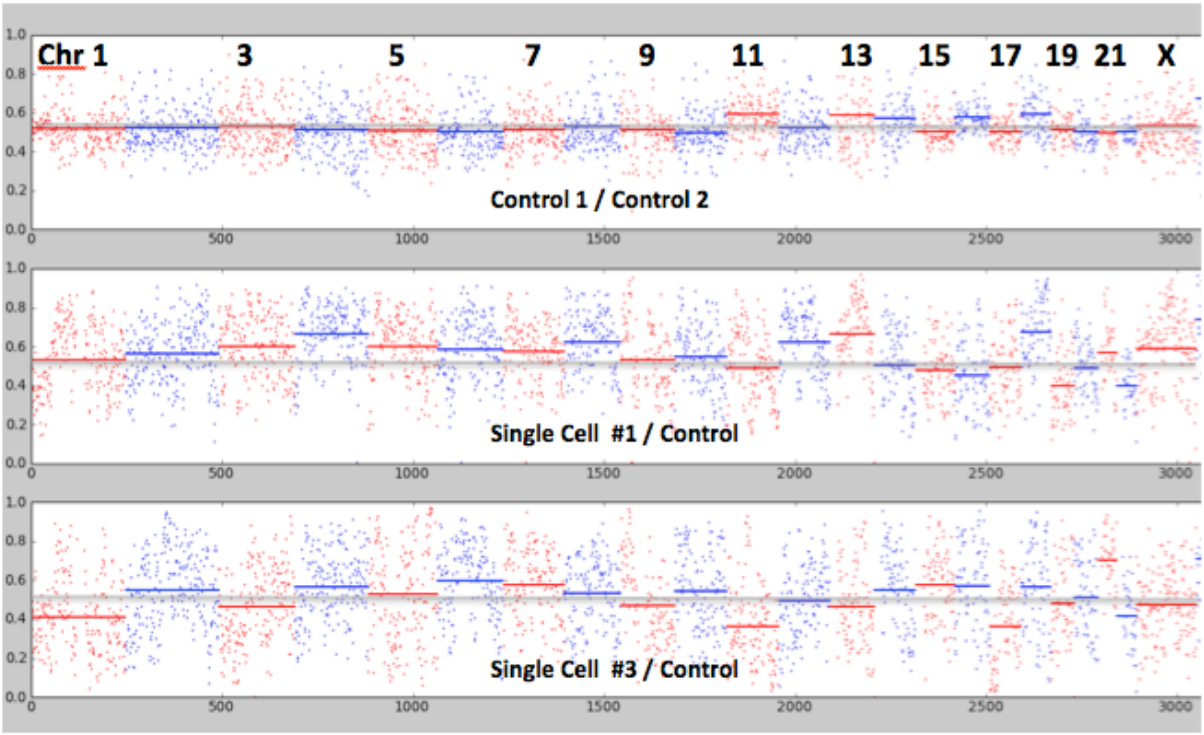


Figure 9.8 The initial test run of single cell DNaseI sensitivity assay. Here shows the density of reads mapped to different chromosomes on the reference genome. The control cells without DNase treatment shows much less variation across the genome and the single cells that were treatment show larger variations. These data might reflect to some degree the chromosome level genome accessibility differences across genome on different cells, but we tend not to draw any conclusion from this data because of the amplification noise.

Unfortunately, due to the noise of single cell whole genome amplification, we were not able to identify the DNase-sensitive region with confidence. The diploid nature of genome makes the problem even more complicated. A particular site that is cut on the paternal strand does not necessarily means the same site is cut on the maternal strand. Therefore, the cutting information is diluted and often masked by the opposite strand, making the data interpretation

much more difficult than would have been with a haploid genome.

In Chapter 8, we introduced a method of digital counting of DNA fragments with resolved haploid structure. The same principle can be used here to count the DNase cutting sites digitally, with resolved allele specific information, as is shown in Figure 9.8.

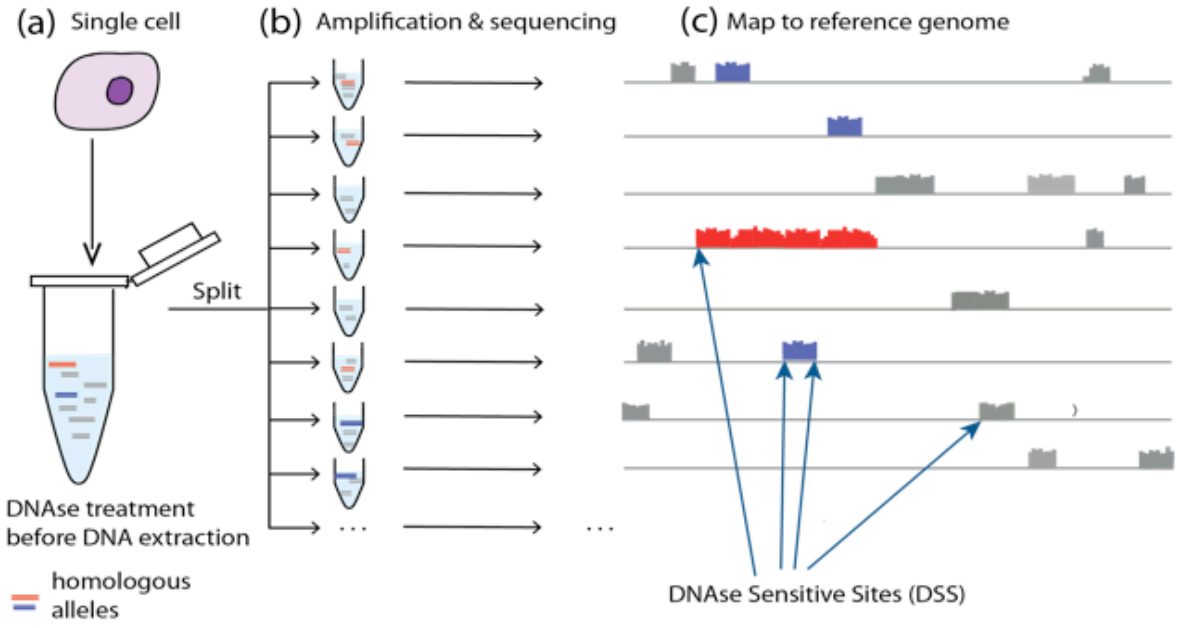


Figure 9.9 Digital counting of DNaseI sensitive sites by digital whole genome amplification (Chapter 8). A single cell is isolated in a reaction tube and treated with DNaseI before lysed to expose its DNA molecules. The DNA molecules are randomly distributed to 24 reaction wells and amplified by MALBAC before barcoded sequenced on an illumina HiSeq 2000 platform. The DNase sensitive sites are then counted digitally by identifying the borders of the fragments from each tube.

By doing this, the fragments are identified independent of the amplification noise. The homolog alleles are also separated which yields allele-specific information on chromatin accessibility. As an example shown in Figure 9.8, the red allele is less sensitive than the blue allele.

As a summary, to study the function of genome at the single cell level, we are trying to study single cell beyond the linear sequence. We developed a new whole-transcriptome amplification method with significantly improved sensitivity and reproducibility. To understand the link between genome and its function (transcriptome), we are exploring the possibility of studying genome accessibility genome-wide at the single cell level.

Reference:

Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard,

P., Mozes, E., Garraway, L.A., and Pe'er, D. (2010). An Integrated Approach to Uncover

Drivers of Cancer. Cell *143*, 1005–1017.

Avner, P., Heard, E., and others (2001). X-chromosome inactivation: counting, choice and initiation. Nature Reviews Genetics *2*, 59–66.

Baker, B.S., Gorman, M., and Marin, I. (1994). Dosage Compensation in Drosophila. Annual Review of Genetics *28*, 491–521.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Bell, O., Tiwari, V.K., Thomä, N.H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. Nature Reviews Genetics *12*, 554–564.

Berger, S.L. (2002). Histone modifications in transcriptional regulation. Current Opinion in Genetics & Development *12*, 142–148.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Blake, W.J., KÆrn, M., Cantor, C.R., and Collins, J.J. (2003). Noise in eukaryotic gene expression. Nature *422*, 633–637.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell *132*, 311–322.

Chiang, M.-K., and Melton, D.A. (2003). Single-Cell Transcript Analysis of Pancreas Development. Developmental Cell *4*, 383–393.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P.S., Rothenberg, M.E., Leyrat, A.A., Sim, S., Okamoto, J., Johnston, D.M., Qian, D., et al. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nature Biotechnology *29*, 1120–1127.

Dean, F.B., Nelson, J.R., Giesler, T.L., and Lasken, R.S. (2001). Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle

Amplification. Genome Res *11*, 1095–1099.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.

Downie, S.E., Flaherty, S.P., Swann, N.J., and Matthews, C.D. (1997). Estimation of Aneuploidy for Chromosomes 3, 7, 16, X and Y in Spermatozoa from 10 Normospermic Men Using Fluorescence in-Situ Hybridization. Mol. Hum. Reprod. *3*, 815–819.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. PNAS *89*, 1827–1831.

Gilbert, S.F. (2010). Developmental Biology, Ninth Edition (Sinauer Associates, Inc.).

Gore, A., Li, Z., Fung, H.-L., Young, J.E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E., et al. (2011). Somatic coding mutations in human induced pluripotent stem cells. Nature *471*, 63–67.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Grunau, C., Clark, S.J., and Rosenthal, A. (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucl. Acids Res. *29*, e65–e65.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature *464*, 1071–1076.

Habermann, J.K., Bündgen, N.K., Gemoll, T., Hautaniemi, S., Lundgren, C., Wangsa, D., Doering, J., Bruch, H.-P., Nordstroem, B., Roblick, U.J., et al. (2011). Genomic instability influences the transcriptome and proteome in endometrial cancer subtypes. Molecular Cancer *10*, 132.

Hall, J.G. (1988). Review and hypotheses: somatic mosaicism: observations related to clinical genetics. Am J Hum Genet *43*, 355–363.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of Cancer: The Next Generation. Cell *144*, 646–674.

Heller, M.J. (2002). DNA MICROARRAY TECHNOLOGY: Devices, Systems, and

Applications. Annual Review of Biomedical Engineering *4*, 129–153.

Hussein, S.M., Batada, N.N., Vuoristo, S., Ching, R.W., Autio, R., Närvä, E., Ng, S., Sourour, M., Hämäläinen, R., Olsson, C., et al. (2011). Copy number variation and selection during reprogramming to pluripotency. Nature *471*, 58–62.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. *21*, 1160–1167.

Keller, P.J., Schmidt, A.D., Wittbrodt, J., and Stelzer, E.H.K. (2008). Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy. Science *322*, 1065–1069.

Klein, C.A., Seidl, S., Petat-Dutter, K., Offner, S., Geigl, J.B., Schmidt-Kittler, O., Wendler, N., Passlick, B., Huber, R.M., Schlimok, G., et al. (2002). Combined transcriptome and genome analysis of single micrometastatic cells. Nat. Biotechnol. *20*, 387–392.

Kurimoto, K., Yabuta, Y., Ohinata, Y., Ono, Y., Uno, K.D., Yamada, R.G., Ueda, H.R., and Saitou, M. (2006). An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucl. Acids Res. *34*, e42–e42.

Lasken, R.S. (2007). Single-cell genomic sequencing using Multiple Displacement Amplification. Current Opinion in Microbiology *10*, 510–516.

Lefort, N., Feyeux, M., Bas, C., Féraud, O., Bennaceur-Griscelli, A., Tachdjian, G., Peschanski, M., and Perrier, A.L. (2008). Human embryonic stem cells reveal recurrent genomic instability at 20q11.21. Nat. Biotechnol. *26*, 1364–1366.

Lefort, N., Perrier, A.L., Laâbi, Y., Varela, C., and Peschanski, M. (2009). Human embryonic stem cells and genomic instability. Regen Med *4*, 899–909.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. , Published Online: 25 December 1993; | Doi:10.1038/366362a0 *366*, 362–365.

Metzenberg, S. (2007). Working With DNA (THE BASICS (Taylor & Francis).

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nature Reviews Genetics *11*, 31–46.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Morozova, O., Hirst, M., and Marra, M.A. (2009). Applications of New Sequencing Technologies for Transcriptome Analysis. Annual Review of Genomics and Human Genetics *10*, 135–151.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., et al. (2011). Tumour evolution inferred by single-cell sequencing. Nature *472*, 90–94.

Nelson, D.L., and Cox, M.M. (2008). Lehninger Principles of Biochemistry (W. H. Freeman).

Rehen, S.K., Yung, Y.C., McCreight, M.P., Kaushal, D., Yang, A.H., Almeida, B.S.V., Kingsbury, M.A., Cabral, K.M.S., McConnell, M.J., Anliker, B., et al. (2005). Constitutional Aneuploidy in the Normal Human Brain. J. Neurosci. *25*, 2176–2180.

Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., et al. (2006). Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nature Methods *3*, 511–518.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science *239*, 487–491.

Shiroguchi, K., Jia, T.Z., Sims, P.A., and Xie, X.S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc. Natl. Acad. Sci. U.S.A. *109*, 1347–1352.

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature *484*, 339–344.

Strahl, B.D., Allis, C.D., and others (2000). The language of covalent histone modifications. Nature *403*, 41.

Szyf, M., Pakneshan, P., and Rabbani, S.A. (2004). DNA demethylation and cancer: therapeutic implications. Cancer Letters *211*, 133–143.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods *6*, 377–382.

Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Science *329*, 533–538.

Tietjen, I., Rihel, J.M., Cao, Y., Koentges, G., Zakhary, L., and Dulac, C. (2003). Single-Cell Transcriptional Analysis of Neuronal Progenitors. Neuron *38*, 161–175.

Wang, J., Fan, H.C., Behr, B., and Quake, S.R. (2012). Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. Cell *150*, 402–412.

Williams, B.R., Prabhu, V.R., Hunter, K.E., Glazier, C.M., Whittaker, C.A., Housman, D.E., and Amon, A. (2008). Aneuploidy Affects Proliferation and Spontaneous Immortalization in Mammalian Cells. Science *322*, 703–709.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R.H., Eshleman, J.R., Nowak, M.A., et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature *467*, 1114–1117.

Youssoufian, H., and Pyeritz, R.E. (2002). Mechanisms and consequences of somatic mosaicism in humans. Nature Reviews Genetics *3*, 748–758.

Yurov, Y.B., Iourov, I.Y., Vorsanova, S.G., Liehr, T., Kolotii, A.D., Kutsev, S.I., Pellestor, F., Beresheva, A.K., Demidova, I.A., Kravets, V.S., et al. (2007). Aneuploidy and Confined Chromosomal Mosaicism in the Developing Human Brain. PLoS ONE *2*, e558.

(2010). The Nucleus (Cold Spring Harbor Laboratory Press).